Saaty

Lectures on Modern Mathematics

Volume III

510
LEC

Wiley

Lectures on Modern Mathematics

Volume III

Edited by T. L. SAATY

Saaty

Lectures on Modern Mathematics

Volume II

510
LEC

Wiley

Saaty

Lectures on Modern Mathematics

Volume III

510
LEC

Wiley

Lectures on Modern Mathematics

edited by **T. L. Saaty**, *formerly at Office of Naval Research presently at Arms Control and Disarmament Agency*

volume III

John Wiley & Sons, Inc., New York • London • Sydney

SECOND PRINTING, JULY, 1967

Copyright © 1965 *by John Wiley & Sons, Inc.*

All Rights Reserved. This book or any part thereof must not be reproduced in any form without the written permission of the publisher, except for any purpose of the United States Government.

Library of Congress Catalog Card Number: 63-20639

Printed in the United States of America

ACCESSION NO. 64818

CLASS NO. 510 LEC

23FEB 76

O/S

Lectures on Modern Mathematics

Γ Δ Θ Λ Ξ Π Σ Φ Ψ Ω

Preface

Jean Dieudonné has advanced the idea* of the good mixture of mathematical spirit (essential to the well being of mathematics) in an individual. He differentiates between the tactician who faces a problem with old and well-tested tools and the strategist who imbeds the problem in a general theory with the result that the final solution often appears as a triviality. Each of the two aspects will suffer without the other, and the two must blend harmoniously as was the case with Hilbert.

With a little bit of tactics and a substantial amount of strategy, the lectures of the third volume of this series again deal with the broad and lively topics of modern mathematics. What has been covered in these three volumes is a modest fraction of the universe of mathematics, a subject that must continue its position in the front ranks of knowledge with much unity and substantial variety.

In the preface to Volume I, I made the following observations, which have overall significance for the series, and should perhaps be repeated.

In our time the growth of mathematical literature, like that of all the sciences, has burgeoned beyond the point where an individual

* Jean Dieudonné, Recent developments in mathematics, *Amer. Math. Monthly*, 71: 3 (March 1964) 239–248.

finds its content accessible. It has been shown that the number of scientists and the number of published scientific papers have multiplied by ten for every doubling of the general population, and hence that the next decade or two will see the doubling of the volume of all the existing literature produced until now. The man who is not a specialist in a given area of research will find that the sheer volume of published work makes him a stranger, and the proverbial esotericism of mathematics has by now limited the possibility of universalism to the very few.

For the future we are promised the applied marvels of data processing and information storage and retrieval. Nonetheless, it requires more than mechanical assimilation to unify and sift for significance the published knowledge of even a relatively small subfield of mathematics. We are still obliged (and shall be for a long visible future) to look to the mature specialist himself for such evaluation, summarization, and interpretation as a partial answer to the "paper explosion."

This is the third and final volume of the series of lectures jointly sponsored by the George Washington University and the Office of Naval Research, which by intelligent summaries further extends the horizons of accessible modern mathematics. The lectures were begun in the fall of 1962 at the Lisner Auditorium of the George Washington University and continued at monthly intervals.

Our intention, as stated in the preface to Volume I, was to invite each of the eminent men represented here to delineate a substantial research area, to describe it broadly and comprehensively for an audience of mathematicians who are not specialists in that area, and to contribute to this description his individual evaluation of the esthetic and practical aspects of the field, its position in mathematical development as a whole, and its future, as that might be implied in the conjectural exposition of its unsolved problems. The speakers have responded to a difficult challenge, compressing the enormous material at their command into a necessarily limited space, preserving the original spirit of the project in their informal approach, refraining from any deep and intricate excursions which might intimidate the tyro, and at the same time giving in each case the personal flavor of their own involvement in mathematical research.

The outstanding success of the lecture series makes us confident that these books will be of interest to all mathematicians desiring to keep abreast of the major achievements in various fields, as well as

to those in the general scientific public wanting to have a flavor of the rapid and sophisticated development of the "Queen of the Sciences." To the graduate student in mathematics embarking on his research career, it should be both useful and encouraging, providing glimpses of large areas to which he may not have been formally introduced and challenging him with unsolved problems.

The last set of lectures was again organized and efficiently administered by Professor David Nelson, who was aided by Professor Norman Wiegmann. To both colleagues go our thanks and appreciation.

T. L. Saaty
Editor

Washington, D. C.
February 1965

List of Lectures

VOLUME I

P. R. Halmos, *A Glimpse into Hilbert Space*
Laurent Schwartz, *Some Applications of the Theory of Distributions*
A. S. Householder, *Numerical Analysis*
Samuel Eilenberg, *Algebraic Topology*
Irving Kaplansky, *Lie Algebras*
Richard Brauer, *Representations of Finite Groups*

VOLUME II

L. Nirenberg, *Partial Differential Equations with Applications in Geometry*
Marshall Hall, Jr., *Generators and Relations in Groups—The Burnside Problem*
R. H. Bing, *Some Aspects of the Topology of 3-Manifolds Related to the Poincaré Conjecture*
Lars Gårding, *Partial Differential Equations: Problems and Uniformization in Cauchy's Problem*
Lars V. Ahlfors, *Quasiconformal Mappings and Their Applications*
J. Milnor, *Differential Topology*

VOLUME III

Einar Hille, *Topics in Classical Analysis*
H. S. M. Coxeter, *Geometry*
Georg Kreisel, *Mathematical Logic*
Paul Erdös, *Some Recent Advances and Current Problems in Number Theory*
Michel Loève, *On Stochastic Processes*
J. Kampé de Fériet, *Random Integrals of Differential Equations*

Contents

1 Topics in Classical Analysis, 1
Einar Hille

2 Geometry, 58
H. S. M. Coxeter

3 Mathematical Logic, 95
Georg Kreisel

4 Some Recent Advances and Current Problems in Number Theory, 196
Paul Erdös

5 On Stochastic Processes, 245
Michel Loève

6 Random Integrals of Differential Equations, 277
J. Kampé de Fériet

1

Topics in Classical Analysis

Einar Hille

PROLOGUE

If the purpose of these lectures is to describe the wave front of current research in classical analysis, my efforts will fall woefully short of the desired goal. In the time and space at my disposal I can only scratch the surface here and there with big gaps in between. Thus a choice of topics must be made and this is a highly subjective matter. It is said that one man's meat is another man's poison. A partial excuse for my choice of the poison is that some of the most succulent meat has already been consumed by other speakers in this series. Thus partial differential equations as well as complex function theory form substantial chapters of classical analysis and they have been treated expertly by other speakers. From what is left I have finally after much hesitation chosen the following topics:

1. Functional inequalities.
2. Functional equations.
3. Mean values.
4. Transfinite diameters.
5. Potential theory.

Obviously the first two topics are closely related and the last three also have much in common. Moreover, mean values as presented here center around a particular class of functional equations.

Various types of functional inequalities arise in many connections. Convex and subadditive functions are important concepts based on such inequalities and so are subharmonic functions (not treated here). Much of the material presented is suggested by uniqueness theorems in the theory of ordinary differential equations. Classical complex function theory and renewal theory form other rich sources. There is no discussion of inequalities in the usual sense. For this topic we refer to the excellent tract by Beckenbach and Bellman [7] in the references at the end of this article. See also the monograph of Kazarinoff [86].

If functional inequalities are products of modern times, functional equations on the other hand have a long history, going back to Newton. In the present report we omit differential, difference, and integral equations and restrict ourselves to what might be called algebroid functional equations. Here, largely through the efforts of the Hungarian school, new life is sprouting and basic problems are being solved. Some of these are discussed here.

Mean values and averaging processes have proved themselves powerful tools in analysis. Summability and ergodic theory are outstanding illustrations. Our modest aim here is to lay the foundation for the theory of transfinite diameters and to call attention to some elementary consequences of the mean value property. Other applications will suggest themselves to the reader.

Transfinite diameters in their role as capacities are basic for potential theory. The latter is at present a very lively topic and could justify a report of its own. For a general survey we refer to the outstanding monograph of M. Tsuji [78], which covers the literature up to 1958. Aside from necessary background material we shall concentrate on recent work.

FUNCTIONAL INEQUALITIES

1.0. *Orientation.* In this chapter we shall consider various types of inequalities between functions. The first four sections deal

with what we shall call, for the lack of a better name, *restrictive inequalities.*

Let F be a given class of functions f on reals to reals defined, say, on the interval $[0, a]$. Let T be a mapping, not necessarily linear, of F into F and consider the inequality

$$f(t) \leqq T(f)(t), \qquad 0 \leqq t \leqq a. \tag{1.0.1}$$

Such an inequality will hold for a certain subset F_0 of F. If $F_0 = F$ we say that the inequality is *trivial* (with respect to F). If F_0 is a proper subset of F, $F_0 \neq \emptyset$, then (1.0.1) is *restrictive.* In particular, the inequality is *determinative* if it holds for one and only one element of F. Finally, it is *absurd* if $F_0 = \emptyset$.

The four cases are illustrated by

$$f(t) \leqq \tfrac{1}{4}\{3 + [f(t)]^4\}, \qquad f(t) \leqq [f(t)]^2,$$

$$f(t) \leqq K \int_0^t f(s)\, ds, \qquad f(t) \leqq -1 - [f(t)]^2,$$

where F is the set of non-negative continuous functions on $[0, a]$.

Section 1.5 deals with some function theoretical inequalities which are usually of a different nature but are restrictive in the sense that the given inequality implies much stronger inequalities. Section 1.6 deals with inequalities on product spaces, convex and subadditive functions being typical instances.

1.1. *Determinative Inequalities.* Such inequalities underlie the classical uniqueness theorems in the theory of differential and integral equations, and all the inequalities listed below come from this source. Thus the classical Lipschitz condition leads to the following determinative inequality:

Proposition 1. If f is a bounded measurable function on $(0, a)$ and if

$$0 \leqq f(t) \leqq K \int_0^t f(s)\, ds, \qquad 0 \leqq t \leqq a, \tag{1.1.1}$$

then $f \equiv 0$.

There are three different types of proof available for this proposition in the literature, and we shall take the liberty of reminding the

reader of them. Using the *method of iteration*, one substitutes repeatedly the inequality into itself and after n steps one has

$$f(t) \leqq \frac{1}{n!} K^{n+1} \int_0^t (t - s)^n f(s)\, ds,$$

and the conclusion is obvious.

The second alternative is to proceed *step by step*. If θ is fixed, $0 < \theta < 1$, and B is the supremum of f in $(0, \theta/K)$ then

$$0 \leqq B \leqq K \frac{\theta}{K} B = \theta B \qquad \text{and} \qquad B = 0.$$

The same argument then applies to the interval $(\theta/K, 2\theta/K)$ and so on.

Proof by integration is less elementary but lends itself better to generalization. Here we set

$$F(t) \equiv \int_0^t f(s)\, ds.$$

Then $F(0) = 0$ and for almost all t

$$F'(t) = f(t) \leqq KF(t),$$

so that

$$\frac{d}{dt}[F(t)e^{-Kt}] \leqq 0$$

and, since $F(0) = 0$,

$$F(t)e^{-Kt} \leqq 0.$$

This gives $F(t) = 0$ for all t and hence also $f(t) = 0$.

The same type of argument gives the following generalization.

Proposition 2. With the same assumptions, let $K(t) > 0$ *and* $K(t) \in L(0, a)$. *Then*

$$0 \leqq f(t) \leqq \int_0^t K(s) f(s)\, ds \qquad \textit{implies } f \equiv 0. \tag{1.1.2}$$

The uniqueness theorem of M. Nagumo [61] leads to the following determinative inequality.

Proposition 3. If $f \in C(0, a)$, if $f(0) = 0$, and if $f'(0)$ exists and is 0, then

$$0 \leqq f(t) \leqq \int_0^t f(s)\,\frac{ds}{s} \qquad \textit{implies } f \equiv 0. \tag{1.1.3}$$

Since $f(0) = f'(0) = 0$,

$$F(t) \equiv \int_0^t f(s)\,\frac{ds}{s}$$

exists and $F'(0) = 0$. Hence

$$F'(t) = t^{-1}f(t) \leqq t^{-1}F(t),$$

and

$$\frac{d}{dt}\,[t^{-1}F(t)] \leqq 0.$$

Since $t^{-1}F(t)$ is decreasing and tends to zero with t, it follows that $F(t) \leqq 0$ and this requires $F(t) = 0$ and hence also $f(t) \equiv 0$.

A more delicate argument is involved in the uniqueness theorem of W. F. Osgood [64].

Proposition 4. Suppose that $\omega(u)$ is defined for $0 \leqq u$, $\omega(0) = 0$, $\omega(u)$ is continuous and strictly increasing, and

$$\int_\delta^a [\omega(u)]^{-1}\,du \to \infty \qquad \textit{as } \delta \downarrow 0. \tag{1.1.4}$$

If $f(t) \in C[0, a]$, then

$$0 \leqq f(t) \leqq \int_0^t \omega[f(s)]\,ds \qquad \textit{implies } f \equiv 0. \tag{1.1.5}$$

Here we introduce

$$g(t) = \max_{0 \leqq s \leqq t} f(s)$$

and note that $f(t) \leqq g(t)$ while for each t there is a $t_1 \leqq t$ such that $f(t_1) = g(t)$.

We may assume that $g(t) > 0$ for $t > 0$. It is then nondecreasing and so is $\omega[g(t)]$. Then

$$\begin{aligned} f(t) \leqq g(t) = f(t_1) &\leqq \int_0^{t_1} \omega[f(s)]\,ds \\ &\leqq \int_0^t \omega[f(s)]\,ds \leqq \int_0^t \omega[g(s)]\,ds, \end{aligned}$$

so that

$$g(t) \leqq \int_0^t \omega[g(s)]\, ds \equiv G(t),$$

where $G(t)$ is strictly increasing. We have now

$$G'(t) = \omega[g(t)] \leqq \omega[G(t)],$$

so that

$$\frac{G'(t)}{\omega[G(t)]} \leqq 1,$$

and

$$\int_\delta^a \frac{G'(t)\, dt}{\omega[G(t)]} < a.$$

On the other hand, $G(t)$ is strictly increasing and differentiable with a continuous derivative, so if we set $u = G(t)$, we get

$$\int_{G(\delta)}^{G(a)} \frac{du}{\omega(u)} < a.$$

Since $G(\delta) \to 0$ with δ, this is a contradiction and $g(t) = f(t) \equiv 0$.

1.2. *Maximal Elements.* We turn now to inequalities that are restrictive and usually not determinative. Here the natural problem is to determine the subset F_0 of F on which the inequality is valid. In the general case very little seems to be known about this problem and the best we can do is to find some properties of f that are implied by the inequality. In the present section we shall assume F to be partially ordered and try to find the maximal element of F_0 under somewhat restrictive assumptions on the transformation T of (1.0.1).

In the following $C[0, a]$ will be the underlying space and F shall be the positive cone $C^+[0, a]$ of the space. For convenience we include the zero element in C^+. F is partially ordered under the convention that $f_1 \leqq f_2$ if $f_2 - f_1 \in C^+$. Let T be a transformation of $C[0, a]$ into itself which is *order preserving*, that is,

$$f_1 \leqq f_2 \quad \text{implies } T[f_1] \leqq T[f_2]. \tag{1.2.1}$$

Such a transformation is obviously *positive* in the sense that it maps C^+ into itself. We further assume that T is a *contraction.*

C^+ is a complete metric space under the sup norm. To say that T is a contraction implies the existence of a constant k, $0 < k < 1$, such that

$$\|T[f_1] - T[f_2]\| \leqq k\|f_1 - f_2\|. \tag{1.2.2}$$

This means that the classical Banach-Cacciopoli fixed-point theorem for such mappings applies (see also F. F. Bonsall [89]). There is a unique element f_0 of C^+ such that, if f is any element of C^+, then

$$\lim_{n\to\infty} T^n[f] = f_0 \tag{1.2.3}$$

and

$$T[f_0] = f_0. \tag{1.2.4}$$

This gives the following proposition.

Proposition 5. If T is an order preserving contraction on $C^+[0, a]$ to itself of fixed point f_0, then

$$f \leqq T[f] \quad \textit{implies } f \leqq f_0. \tag{1.2.5}$$

For we have

$$f \leqq T[f] \leqq T^2[f] \leqq \cdots \leqq T^n[f] \leqq \lim T^n[f] = f_0.$$

Thus we have found the maximal element of F_0. Ordinarily it is not true that

$$F_0 = [f \,|\, f \leqq f_0],$$

so we are still far from a complete description of F_0. The inequality becomes determinative if and only if f_0 is the zero element. It should be noted that if the first inequality in (1.2.5) is reversed so is the second.

In the theory of differential equations it is customary to convert the equation into an integral equation and to use the latter for a priori estimates of the rate of growth of the solutions. A basic tool is the following proposition which we give essentially in the formulation of Bourbaki [8, p. 10].

Proposition 6. Let f and $g \in C^+[0, a]$, let $K(t) > 0$, and $K(t) \in L(0, a)$. Then the inequality

$$f(t) \leqq g(t) + \int_0^t K(s)f(s)\,ds \tag{1.2.6}$$

implies that

$$f(t) \leqq f_0(t) \equiv g(t) + \int_0^t g(s)K(s) \exp\left[\int_s^t K(u)\,du\right] ds. \tag{1.2.7}$$

If $T[f]$ is defined by the right member of (1.2.6), we see that

$$f_0 = T[f_0]$$

and f_0 is the only fixed point of the transformation. This is a linear transformation on C^+ to itself and hence order preserving. It is a contraction if and only if

$$\int_0^a K(s)\,ds = k < 1.$$

For the general case where this condition is not satisfied, we can use the method of integration sketched above. Even in the simplest case, $K \equiv 1$, it does not seem clear that all of F with $f \leqq f_0$ belongs to F_0.

Propositions 1 and 2 are special cases of 6. Here the inequality is determinative since f_0 is the zero element. The transformations T of Propositions 3 and 4 also have the zero element as only fixed points.

The special case where T is a convolution will be discussed in Section 1.4.

1.3. *Bounds.* Many problems in analysis call for nontrivial lower as well as upper bounds. The following is an example of such a problem where the bounds are obtained by using a class of simultaneously valid restrictive inequalities.

Consider a linear differential equation

$$W'(t) = A(t)W(t), \tag{1.3.1}$$

where A and W are n-by-n matrices, more generally elements of a complex noncommutative Banach algebra. It is supposed that $A(t)$ is continuous for $0 < t \leqq a$ and it is required to find bounds for $W(t)$ as $t \downarrow 0$.

Proposition 7. Under the stated assumptions

$$\|W(a)\| \exp\left[-\int_t^a \|A(s)\|\,ds\right] \leqq \|W(t)\| \leqq \|W(a)\| \exp\left[\int_t^a \|A(s)\|\,ds\right]. \tag{1.3.2}$$

This is obtained from

$$W(\beta) = W(\alpha) + \int_\alpha^\beta A(s)W(s)\,ds,$$

valid for $0 < \alpha < \beta \leqq a$. Setting

$$\|W(s)\| = f(s), \qquad \|A(s)\| = K(s),$$

we obtain

$$f(t) \leqq f(\alpha) + \int_\alpha^t K(s)f(s)\,ds.$$

This is of type (1.2.6) and gives

$$f(\beta) \leqq f(\alpha) \exp\left[\int_\alpha^\beta K(s)\,ds\right].$$

Setting $\alpha = t$, $\beta = a$, we get the first half of (1.3.2). To get the second half we operate with

$$f(t) \leqq f(\beta) + \int_t^\beta K(s)f(s)\,ds$$

and use the method of integration.

The argument extends immediately to the complex plane. In particular we see that $\|W(t)\|$ is bounded away from 0 and ∞, if $\|A(t)\| \in L(0, a)$. Taking

$$A(t) = t^{-1-p}B(t), \qquad B(t) \text{ bounded}, \tag{1.3.3}$$

we get the classical estimates for regular singular points if $p = 0$ and for irregular singular points of rank p if p is a positive integer.

If, instead, $W(t)$ is known to be continuous for $a \leqq t < \infty$, then we replace (1.3.2) by

$$\|W(a)\| \exp\left[-\int_a^t \|A(s)\|\,ds\right] \leqq \|W(t)\| \leqq \|W(a)\| \exp\left[\int_a^t \|A(s)\|\,ds\right]. \tag{1.3.4}$$

If, in particular,

$$\lim_{t\to\infty} \|W(t)\| \exp\left[\int_a^t \|A(s)\|\, ds\right] = 0, \qquad \text{then } W(t) \equiv 0. \tag{1.3.5}$$

The following example belongs to a different range of ideas. Suppose that $P(u)$ is a polynomial in u of even degree and real coefficients such that the equation

$$P(u) - u = 0$$

has exactly two real roots, r_1 and r_2, and that $P(u) - u$ is positive outside the interval (r_1, r_2). Then clearly the inequality

$$f(t) \geqq P[f(t)] \qquad \text{implies } r_1 \leqq f(t) \leqq r_2 \tag{1.3.6}$$

for all t.

A variant of this simple observation was used by Hille [39] in a discussion of the asymptotic behavior of the function $m(\lambda, \alpha)$ introduced by H. Weyl in his discussion of singular boundary value problems. Stripped of unessential features the inequality can be written

$$|F(t)| \geqq |\Im[F(t)]| \geqq t\left[|F(t) - a|^2 - \frac{A}{t}|F(t) - a|\,|F(t) - b| + \frac{B^2}{t^2}|F(t) - b|^2\right]. \tag{1.3.7}$$

Here $F(t)$ is a complex-valued continuous function on $[1, \infty)$, a, b, A, B are real numbers, $A > 0, B > 0, (A - 2B)|b - a| > -1$. From this inequality we conclude that $|F(t)|$ must stay bounded as $t \to \infty$; this being the case, $\lim F(t)$ must exist and equal a. We then set

$$F = a + \frac{G}{t}, \tag{1.3.8}$$

and note that the imaginary part of F equals t^{-1} times the imaginary part of G. This gives a new inequality for $g = |G|$, namely

$$g \geqq g^2 - Ag\left(\gamma + \frac{g}{t}\right) + B^2\left(\gamma - \frac{g}{t}\right)^2, \qquad \gamma = |b - a|. \tag{1.3.9}$$

Here the condition $(A - 2B)\gamma + 1 > 0$ ensures that the corresponding quadratic equations have positive roots bounded away from 0 and ∞. It follows that $G(t)$ is bounded away from 0 and ∞ for all t, which was the desired result.

1.4. *Convolution Inequalities.* We return to inequalities of the type considered in Section 1.2. T is now supposed to be linear, more precisely a convolution.

Let G be a locally compact Abelian additive group and μ a positive regular measure of total mass one on G. Let F be the class of real-valued continuous functions of f on G such that for each x

$$(f * \mu)(x) = \int_G f(x - s)\, d\mu(s) \tag{1.4.1}$$

converges absolutely. Let F_0 be the subclass of F such that on G

$$f - f * \mu \geqq 0. \tag{1.4.2}$$

It will be supposed that μ has no mass at the origin. Let $G(\mu)$ be the closed subgroup of G generated by the elements in the support of μ. G is normally taken to be either R^n or Z^n.

Such inequalities are of some interest to the theory of probability, especially for renewal theory. This aspect of the question has been treated in recent papers by Choquet and Deny [16], W. Feller [28], Feller and Orey [29], Hoeffding [42], Karlin [48], and Spitzer [74]. The results listed below are in the main taken from the Uppsala dissertation of Matts Essén [21] which is devoted to convolution inequalities per se.

In the present case $\|T\| = 1$ so Proposition 5 does not apply and the fixed points of the transformation are neither unique nor of special importance for the problem at hand. The bounded solutions of

$$f = f * \mu \tag{1.4.3}$$

have been found by Choquet and Deny [16]. They are periodic functions whose periods contain the elements of the support of μ. In particular, if $G(\mu) = G$, then any constant is a solution of (1.4.3) and there are no other bounded solutions. There may

be unbounded solutions, however. Thus if

$$(1.4.4)\quad G = G(\mu) = R^1, \quad \int_{-\infty}^{\infty} |x|\, d\mu(x) < \infty, \quad \int_{-\infty}^{\infty} x\, d\mu(x) = 0,$$

then any function of the form $f = Ax + B$, where A, B are arbitrary constants, is a solution of (1.4.3). Such solutions are referred to as trivial solutions of (1.4.2) in the following.

Let $G = R^n$ or Z^n and assume that $G(\mu) = G$.

Proposition 8. If $n = 1$ and if $\int_G |x|\, d\mu(x) < \infty$, then (1.4.2) *has a nontrivial bounded solution if and only if $\int_G x\, d\mu(x) \neq 0$.*

If $n = 2$ and if $\int_G |x|^{1+\alpha}\, d\mu(x) < \infty$ for some positive α, then a sufficient condition for the existence of solutions is the existence of a linear function $l(x) = ax^1 + bx^2$ such that

$$\int_G l(x)\, d\mu(x) \neq 0$$

and this condition is also necessary if $\alpha = 1$ is admissible.

If $n \geqq 3$ nontrivial solutions always exist.

Essén proves the following properties of his solutions.

Proposition 9. Let $G = G(\mu) = R^1$ and assume that

$$(1.4.5)\quad \int_{-\infty}^{\infty} |x|\, d\mu(x) < \infty, \qquad \int_{-\infty}^{\infty} x\, d\mu(x) = m \neq 0.$$

If a bounded function f satisfies (1.4.2) *then*

$$(1.4.6)\qquad f - f * \mu \in L^1(R^1).$$

If f is also slowly decreasing, then $f(x)$ tends to limits as $x \to \pm\infty$ and

$$(1.4.7)\qquad f(\infty) - f(-\infty) = \frac{1}{m} \int_{-\infty}^{\infty} [f - f * \mu](x)\, dx.$$

If the left member is zero, then f satisfies (1.4.3).

These results may be sharpened if μ satisfies more restrictive conditions. We state the results for $n = 1$, $G = R^1$.

Proposition 10. Let $G = G(\mu) = R^1$ *and let there be an* α, $0 \leqq \alpha \leqq 1$, *such that*

$$\int_{-\infty}^{\infty} |x|^{1+\alpha}\, d\mu(x) < \infty,$$

while

$$\int_{-\infty}^{\infty} x\, d\mu(x) = 0.$$

Let f *be a solution of* (1.4.2). *If* $f(x) = O(|x|^{\alpha})$, *then* $f - f * \mu \in L^1(R^1)$. *If instead* $f(x) = o(|x|^{\alpha})$, *then* f *satisfies* (1.4.3). *If* $\alpha = 1$ *is admissible and if* $f(x)/x$ *is bounded for* $|x| > 1$ *and slowly oscillating, then* $f(x)/x$ *tends to finite limits when* $x \to \pm\infty$ *and*

$$\lim_{x\to\infty} \frac{f(x) + f(-x)}{x} = -\frac{2}{\sigma^2}\int_{-\infty}^{\infty} [f - f * \mu](x)\, dx, \tag{1.4.8}$$

$$\sigma^2 = \int_{-\infty}^{\infty} x^2\, d\mu(x).$$

Here the terms "slowly decreasing" and "slowly oscillating" have their usual significance in the theory of Tauberian theorems, and the proofs of limit results are based on the Tauberian theorem of Norbert Wiener [83].

1.5. *Some Function Theoretical Inequalities.* Analytic function theory presents many inequalities which are restrictive in the sense that they imply stronger inequalities which must be satisfied by the function. Elsewhere [37, I, p. 200] I have called attention to the principle of moderation which underlies the behavior of analytic functions. It implies that a strong local constraint imposed on such a function either becomes a global constraint or the function must compensate by being strongly unbounded.

The principle of the maximum is a case in point: If $f(z)$ is holomorphic in a bounded domain D and if

$$\limsup |f(z)| \leqq M \tag{1.5.1}$$

whenever z approaches a point of ∂D, then $|f(z)| \leqq M$ everywhere in D. The condition becomes determinative if we add the constraint

$$f(z_0) = Me^{i\alpha},$$

α real, $z_0 \in D$, for then $f(z) \equiv Me^{i\alpha}$.

Here there is no alternative. But if we drop the assumption that D is bounded and assume that (1.5.1) holds merely for approach to finite boundary points, then either $|f(z)| \leqq M$ in D or $f(z)$ is unbounded.

The theorem of Liouville may also be regarded as a restrictive inequality: If $f(z)$ is entire and if

$$M(r;f) = \max_{\theta} |f(re^{i\theta})| = o(r^{n+1}), \tag{1.5.2}$$

then $f(z)$ is a polynomial of degree $\leqq n$. Here we can replace (1.5.2) for instance by

$$M_2(r;f) = \left[\frac{1}{2\pi}\int_0^{2\pi} |f(re^{i\theta})|^2\, d\theta\right]^{1/2} = o(r^{n+1}). \tag{1.5.3}$$

A deeper result is that due to M. H. Stone [75] who showed that the same conclusion is valid if there exists an non-negative trigonometric polynomial $P(\theta)$ such that

$$P(\theta)|f(re^{i\theta})| = o(r^{n+1}), \tag{1.5.4}$$

uniformly in θ. For a proof see also Hille and Phillips [41].

It is not our intention to make an inventory of restrictive inequalities in complex function theory; however, we shall mention some which govern the behavior of functions holomorphic in the right half-plane. Suppose that

$$\limsup_{x \downarrow 0} |f(x + iy)| \leqq 1 \tag{1.5.5}$$

for all y. Under what further restrictions can we conclude that $|f(z)| \leqq 1$ in the right-half-plane? In 1904 Phragmén [67] showed that

$$\lim_{r \to \infty} r^{-1} \log M(r;f) = 0 \tag{1.5.6}$$

is a necessary and sufficient condition. This was generalized four years later by Phragmén in collaboration with Lindelöf [68], but we shall not digress on these results. We proceed to the next mile-

stone, the work of F. and R. Nevanlinna [63]. They introduced the function

$$m(r;f) \equiv \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} \log^+ |f(re^{i\theta})| \cos\theta \, d\theta \tag{1.5.7}$$

and proved

Proposition 11. If (1.5.6) *holds, then either* $|f(z)| \leqq 1$ *or*

$$\liminf_{r\to\infty} r^{-1} m(r;f) > 0. \tag{1.5.8}$$

This was in 1922. In 1937 Ahlfors [4] proved that $r^{-1}m(r;f)$ is nondecreasing so that "lim inf" can be replaced by "lim." This limit may be $+\infty$ however. In 1946 M. Heins [35] gave a simpler proof, and he also proved the two following propositions.

Proposition 12. With $0 \leqq \lambda \leqq \infty$

$$\lim r^{-1} \log M(r;f) = \lambda \tag{1.5.9}$$

exists. If $0 < \lambda < \infty$, *then either*

$$\log M(r;f) < \lambda r \tag{1.5.10}$$

for all r *or* $f(z) = Ce^{\lambda z}$ *with* $|C| = 1$.

Proposition 13. The limits

$$\lim r^{-1} \log M(r;f) = \lambda, \qquad \lim r^{-1} m(r;f) = \nu \tag{1.5.11}$$

are cofinite and if finite then

$$\nu = \tfrac{1}{2}\pi\lambda. \tag{1.5.12}$$

We also mention another result of Heins [36], later extended by B. Kjellberg [49] in whose formulation it reads:

Proposition 14. Let $f(z)$ *be a nonconstant entire function and let* $0 < \alpha < 1$. *Then either the logarithm of the minimum modulus exceeds* $\cos \alpha\pi \log M(r;f)$ *for arbitrarily large values of* r *or*

$$g_\alpha(r) = r^{-\alpha} \log M(r;f) \tag{1.5.13}$$

tends to a positive limit or to $+\infty$ *when* $r \to \infty$.

The case $\alpha = \frac{1}{2}$ was that considered by Heins. The point of main interest to us here is a convolution inequality used by Kjellberg in his proof. He found a positive kernel $K_\alpha(s, r)$ with

$$\int_0^\infty K_\alpha(s, r)\, ds \equiv 1$$

such that

$$g_\alpha(r) < \int_0^\infty g_\alpha(s) K_\alpha(s, r)\, ds \tag{1.5.14}$$

and from this he could conclude the existence of $\lim g_\alpha(r)$.

With a suitable change of variables, this inequality can be reduced to the type considered by Matts Essén (see Section 1.4) and Kjellberg's theorem becomes a consequence of Proposition 9. This was pointed out by Essén in [22].

1.6. *Functional Inequalities on Product Spaces.* We shall be concerned with some inequalities of the general type

$$f[F(x, y)] \leqq H[f(x), f(y)], \tag{1.6.1}$$

where F and H are given functions. Here H is a function on $R^1 \times R^1$ to R^1. In general, x and y are points in a locally compact space S and f is a mapping on S to R^1 while F is a mapping of $S \times S$ to R^1. In the most important applications, S is either R^1 or R_+^1, the set of positive numbers. We shall consider some particular cases.

I. *Convex functions.* Here $S = R^1$,

$$F(x, y) = \tfrac{1}{2}(x + y), \qquad H(u, v) = \tfrac{1}{2}(u + v),$$

so the inequality becomes

$$f\left(\frac{x + y}{2}\right) \leqq \frac{1}{2}[f(x) + f(y)]. \tag{1.6.2}$$

This inequality, first considered by O. Hölder [44] and in more detail by Jensen [46] in 1906, characterizes convex functions. Geometrically, (1.6.2) means that the middle point of any chord of the curve $y = f(x)$ lies above or on the curve. If f is bounded above and convex in some interval (a, b), then f is continuous in (a, b) and has

continuous left- and right-hand derivatives everywhere and these derivatives are increasing functions of x. If f is twice differentiable, then $f''(z) \geqq 0$ is necessary and sufficient for f to be convex. We refer the reader to Hardy, Littlewood, and Pólya [34] for further information.

II. *Subadditive functions.* Here $S = R^1$ or R_+^1,

$$F(x, y) = x + y, \qquad H(u, v) = u + v,$$

and the inequality is

$$f(x + y) \leqq f(x) + f(y). \tag{1.6.3}$$

The case where f is also positive-homogeneous goes back to H. Minkowski [58] in 1896; the general theory was developed by Hille and extensions to R^n by R. A. Rosenbaum [72]. If f is measurable and different from $+\infty$, then f is bounded above on any compact set K of S. If f is also different from $-\infty$, then f is bounded on any K. Any decreasing function satisfies (1.6.3) in R_+^1. It follows that if $S = R_+^1$, then f can tend arbitrarily fast to $+\infty$ as $x \downarrow 0$ and arbitrarily fast to $-\infty$ as $x \to \infty$. If $S = R^1$ and f is finite, then

$$-\infty < \alpha = \lim_{x \to -\infty} \frac{f(x)}{x} = \sup_{x<0} \frac{f(x)}{x} \leqq \inf_{x>0} \frac{f(x)}{x} = \lim_{x \to +\infty} \frac{f(x)}{x} = \beta < \infty. \tag{1.6.4}$$

Subadditive functions play a basic role in the analytical theory of semigroups. For further details see Hille and Phillips [41]. We add that any modulus of continuity is subadditive.

C. T. Ionescu Tulcea [45] has considered an important generalization of (1.6.3), namely

$$f(x + y) \leqq g(x) + h(y). \tag{1.6.5}$$

If f, g, h cannot take the value $+\infty$ and if h is measurable, then f is bounded above on compact sets.

III. *Suboperative functions.* Here S is arbitrary, $H = u + v$, and $F(x, y) = x \circ y$, an associative composition defined on $S \times S$.

Thus the inequality is

$$f(x \circ y) \leqq f(x) + f(y). \tag{1.6.6}$$

A particular case occurs in Hille-Phillips [41]; the general theory with many extensions is due to C. T. Ionescu Tulcea [45] who also considered the inequality

$$f(x \circ y) \leqq g(x) + h(y). \tag{1.6.7}$$

These functions play a basic role in the theory of Lie semigroups of operators.

IV. *Non-Archimedean valuations.* S is in general a field or a ring, but we shall restrict ourselves to $S = R^1$, $R_+{}^1$, or $Z_+{}^1$, where $Z_+{}^1$ denotes the set of positive integers. Here

$$F(x, y) = x + y, \qquad H(u, v) = \max(u, v),$$

and the inequality is

$$f(x + y) \leqq \max[f(x), f(y)]. \tag{1.6.8}$$

If S is $R_+{}^1$ or $Z_+{}^1$, then any decreasing function of x satisfies (1.6.8). This means that in $R_+{}^1$ the function f may tend arbitrarily fast to $+\infty$ when $x \downarrow 0$ and arbitrarily fast to $-\infty$ when $x \to +\infty$. It is bounded above on compact subsets of R^1 if $f \not\equiv +\infty$.

Besides its basic role in valuation theory (see, for instance, Albert [5]), this inequality plays a role for the so-called Čebyšev constant. See Section 4.2 of this chapter.

There are obvious extensions to inequalities of the form

$$f(x \circ y) \leqq \max[f(x), f(y)] \tag{1.6.9}$$

which may not have been explored.

FUNCTIONAL EQUATIONS

2.0. *Orientation.* Functional equations have fascinated mathematicians at least since the days of Euler who encountered such equations when he introduced homogeneous functions. They have a tendency of arising in the most varied connections: parallelogram

of forces (d'Alembert, Poisson, Cauchy), associativity (Abel), non-Euclidean geometry (Bolyai and Lobatčevski), optics (Stokes), determinants and addition theorems (Weierstrass), and so on.

What is to be meant by a functional equation is still debatable even if we agree to exclude differential, difference, and integral equations. The classification problem presents many difficulties. It used to be said that every functional equation requires its own mode of attack. In recent years the situation has improved. Gradually more general results are available, the classically known solutions are shown to be valid under less severe restrictions, existence proofs applicable to a wide range of equations are being found, and uniqueness questions can be settled.

For a general account of the theory and a number of its applications, the reader is referred to the excellent monograph by J. Aczél [2]. The following pages are intended to call attention to some recent results. Again the choice of material is arbitrary and reflects the author's interests. Much of it is inspired by the conferences on functional equations held in Oberwolfach in 1962 and 1963.

There are four sections: *Measurability and Continuity*, *A General Method*, *Uniqueness Theorems*, and *Miscellanea*.

2.1. *Measurability and Continuity.* In the early work on functional equations the assumption was usually made, implicitly or explicitly, that the solution be a continuous function. Thus, in his discussion of the functional equation

$$f(x + y) = f(x) + f(y), \tag{2.1.1}$$

Cauchy [13] in 1821 assumed f to be continuous everywhere and could then show that

$$f(x) = cx,$$

where c is an arbitrary constant. Under the same assumption he could also solve the other three equations which are associated with his name:

$$f(xy) = f(x) + f(y), \qquad f(x + y) = f(x)f(y), \qquad f(xy) = f(x)f(y). \tag{2.1.2}$$

Darboux [17] in 1875 showed for (2.1.1) that continuity everywhere could be replaced by continuity at a single point. Gradually other

conditions were found: onesided boundedness in an interval and finally in a set of positive Lebesgue measure. The latter condition is essentially necessary since Hamel [33] in 1905 succeeded in showing the existence of nonmeasurable solutions bounded in no set of positive measure.

If we assume f to be integrable Lebesgue, say in the interval $(0, \omega)$, then (2.1.1) gives

$$\begin{aligned} f(x) &= \int_0^1 f(x+y)\,dy - \int_0^1 f(y)\,dy \\ &= \int_x^{x+1} f(s)\,ds - \int_0^1 f(s)\,ds, \end{aligned} \tag{2.1.3}$$

so that f is necessarily continuous and hence also differentiable. Thus Cauchy's solution is the only integrable solution.

This is a fairly typical situation: *measurability implies continuity.* Among recent results along such lines let us quote one due to C. T. Ionescu Tulcea [45].

We need some definitions and notation. Let S be a locally compact space on which is defined a positive Radon measure μ satisfying certain mild restrictions. There is a composition law $(s, t) \to s \circ t$ defined for $(s, t) \in D \subset S \times S$ and this mapping is continuous in D. Let S_f, S_g, S_h be three Hausdorff spaces and let G be a mapping of $S_g \times S_h$ into S_f, continuous on every compact set $K \subset S_g \times S_h$.

Proposition 15. Let f, g, h be three mappings of S into S_f, S_g, S_h, respectively. Suppose g and h are μ-measurable and that for every $(s, t) \in D$

$$f(s \circ t) = G[g(s), h(s)]. \tag{2.1.4}$$

Then f is continuous on S.

This theorem applies to a large number of functional equations commonly considered.

We saw earlier that any continuous solution of (2.1.1) is also differentiable and it obviously admits of derivatives of all orders. This does not imply, however, that in the complex plane a continuous solution of

$$f(z_1 + z_2) = f(z_1) + f(z_2)$$

is necessarily analytic. Indeed

$$f(x + iy) = ax + by$$

is a continuous solution and this is analytic if and only if $b = ia$. The same phenomenon holds for other equations of the type

$$f(z_1 \circ z_2) = G[f(z_1), f(z_2)]. \tag{2.1.5}$$

On the other hand, if $G(u, v)$ is a symmetric analytic function of u, v and if the rule of composition is symmetric and analytic, then analytic solutions may be expected. This does not seem to have been proved in full generality, however.

Let us add some remarks pertaining to the case

$$f(x + y) = G[f(x), f(y)] \tag{2.1.6}$$

with a symmetric function G, holomorphic in some domain $D = D_0 \times D_0$ in $C^2 = C \times C$. Let a be a root of the equation

$$G(a, a) = a, \qquad a \in D_0. \tag{2.1.7}$$

We exclude the case in which this equation is an identity or has no solutions in D_0. Equation (2.1.6) has constant solutions, $f(x) \equiv a$, where a is any root of (2.1.7). Suppose now that $f(x)$ is a solution of (2.1.6), defined and continuous in some interval $[0, \omega]$ with $f(0) = a$, $f(x) \not\equiv a$. Suppose that $f(x) \in D_0$ and that

$$G_u'(a, a) = 1. \tag{2.1.8}$$

Then it may be shown that $f(x)$ has derivatives of all orders in $[0, \omega]$ and the higher right-hand derivatives at $x = 0$ are uniquely determined by $f(0)$ and $f'(0)$. This follows from results due to Dunford and Hille [19] where the theorem is proved assuming that $f(x)$ has values in a Banach algebra. It follows that if f and g are two continuous solutions of (2.1.6) and if $f(0) = g(0)$, $f'(0) = g'(0)$, then f and g are identical in $[0, \omega]$. This does not exclude the possibility that these solutions may be extended to the complex plane in two different ways, both extensions satisfying (2.1.6). At most one of these extensions can be analytic.

The case where G is algebraic and $x \circ y = x + y$ has been examined in great detail. Here the analytic solutions are known since

Weierstrass. Continuous solutions of a real variable have been studied by J. F. Ritt [71], P. Montel [59], and P. J. Myrberg [60]. Here it is found that any continuous solution is piecewise analytic. If $f(x)$ is a continuous solution, then there is an analytic function $g(z)$ which is algebraic in z or e^{az} or $\wp(bz)$ and constants α_k such that in the kth interval $f(x) = g(x + \alpha_k)$. Myrberg has indicated that the results extend to associative composition rules.

2.2. *A General Method.* E. Vincze [80] has introduced a method of solving functional equations applicable to a rather large class of equations. It is based essentially on the notion of linear dependence and its expression in terms of identities involving determinants. Let

$$\Delta[F_1(z_1), F_2(z_2), \ldots, F_n(z_n)] = \Delta[F_1, F_2, \ldots, F_n] \tag{2.2.1}$$

denote the determinant

$$\begin{vmatrix} F_1(z_1) & F_2(z_1) & \cdots & F_n(z_1) \\ F_1(z_2) & F_2(z_2) & \cdots & F_n(z_2) \\ \cdots & \cdots & \cdots & \cdots \\ F_1(z_n) & F_2(z_n) & \cdots & F_n(z_n) \end{vmatrix}.$$

Here the functions are defined in some arbitrary set S of the complex plane. Suppose that the $n(k+1)$ functions $F_1(z), \ldots, F_{kn+n}(z)$ satisfy the identity

$$\sum_{\nu=0}^{k} \Delta[F_{\nu n+1}(z_1), F_{\nu n+2}(z_2), \ldots, F_{\nu n+n}(z_n)] = 0 \tag{2.2.2}$$

for all $z_1, z_2, \ldots, z_n$ in S. Let $F_0(z)$ be an arbitrary function. Then the equation

$$\sum_{\nu=0}^{k} \Delta[F_{\nu n+1}(z_1), F_{\nu n+2}(z_2), \ldots, F_{\nu n+n}(z_n), F_0(z_{n+1})] = 0 \tag{2.2.3}$$

also holds for all values of $z_1, z_2 \ldots, z_n, z_{n+1}$ in S. Vincze calls the passage from the first equation to the second "extension by $F_0(z)$." It is the basic tool in his discussion.

Let us further note that $F_1, F_2, \ldots, F_n$ are linearly independent

in S if and only if

$$\Delta[F_1(z_1), F_2(z_2), \ldots, F_n(z_n)] = 0 \tag{2.2.4}$$

for all $z_1, z_2, \ldots, z_n$ in S.

Vincze illustrates his method on a number of examples. The following is one of the simplest. It is desired to find functions F, G, H, K, L satisfying the equation

$$F(z_1 \circ z_2) = G(z_1) + H(z_2) + K(z_1)L(z_2). \tag{2.2.5}$$

Here $z_1, z_2, z_1 \circ z_2$ belong to a set S which is an Abelian semigroup with respect to the operation "$\circ$." In particular, it is assumed that there is a point $a \in S$ such that the equation $a \circ z = c$ has at least one solution for every given c.

Since the left side of (2.2.5) is symmetric in z_1 and z_2,

$$G(z_1) + H(z_2) + K(z_1)L(z_2) = G(z_2) + H(z_1) + K(z_2)L(z_1)$$

or, in determinant notation,

$$\Delta(G, 1) + \Delta(1, H) + \Delta(K, L) = 0. \tag{2.2.6}$$

Here we "extend by 1" and obtain

$$\Delta(G, 1, 1) + \Delta(1, H, 1) + \Delta(K, L, 1) = 0.$$

Since the first two determinants are zero, we are left with

$$\Delta(K, L, 1) = 0, \tag{2.2.7}$$

that is, the functions K, L, 1 are linearly dependent. There are two possibilities, either

$$\text{A: } L(z) \equiv \beta_1, \qquad \text{or B: } K(z) = \beta_1 + \beta_2 L(z). \tag{2.2.8}$$

We shall carry through the discussion of case A. Substituting in (2.2.6) we obtain

$$\begin{aligned} &\Delta(G, 1) + \Delta(1, H) + \Delta(K, \beta_1) \\ &\quad = \Delta(G, 1) - \Delta(H, 1) + \Delta(\beta_1 K, 1) \\ &\quad = \Delta(G - H + \beta_1 K, 1) = 0. \end{aligned}$$

Thus there is a constant β_2 such that

$$G(z) - H(z) + \beta_1 K(z) = 2\beta_2.$$

Combining we get

$$(2.2.9) \quad F(z_1 \circ z_2) = [G(z_1) + \beta_1 K(z_1) - \beta_2] + [G(z_2) + \beta_1 K(z_2) - \beta_2].$$

This is a so-called Pexider equation after H. W. Pexider [66], who considered such equations in 1903.

The equation

$$(2.2.10) \qquad f(z_1 \circ z_2) = f(z_1) + f(z_2)$$

is a generalized Cauchy equation and reduces to one of the classical equations if the operation $\circ$ is either addition or multiplication. Suppose that we can solve (2.2.10) and that $f(z)$ is a solution. Then we can also solve (2.2.9) by setting

$$F(z) = f(z) + 2\beta_3, \qquad G(z) = f(z) - \beta_1 K(z) + \beta_2 + \beta_3.$$

Thus the final solution in case A is

$$F(z) = f(z) + 2\beta_3, \qquad G(z) = f(z) - \beta_1 K(z) + \beta_2 + \beta_3,$$
$$H(z) = f(z) + \beta_3 - \beta_2, \qquad K(z) \text{ arbitrary}, L(z) \equiv \beta_1.$$

Case B is more complicated and involves also the solution of another generalized Cauchy equation

$$(2.2.11) \qquad h(z_1 \circ z_2) = h(z_1)h(z_2).$$

Aside from this no new principle is involved.

Vincze has also applied his method to generalized sine and cosine equations where addition is replaced by an associative operation $\circ$. The method appears to be quite powerful.

2.3. *Uniqueness Theorems.* A simple uniqueness theorem has recently been found by Aczél [3].

Proposition 16. Let $H(u, v, x, y)$ be strictly monotone with respect to u (or v). Let F be continuous for all (x, y) in $\langle A, B\rangle \times \langle A, B\rangle$

where $\langle A, B\rangle$ *is an arbitrary interval. Let*

$$x < F(x, y) < y \qquad \text{if } A < x < y < B. \tag{2.3.1}$$

Then the functional equation

$$f[F(x, y)] = H[f(x), f(y), x, y] \tag{2.3.2}$$

has at most one continuous solution defined in $\langle A, B\rangle$ *and satisfying the initial conditions*

$$f(a) = c, \qquad f(b) = d, \qquad A < a < b < B. \tag{2.3.3}$$

The proof is so simple and elegant that we shall sketch it here. Suppose that there are two distinct continuous solutions f_1 and f_2 and consider first the interval $[a, b]$. If there is an $x_1 \in (a, b)$ where $f_1(x_1) \neq f_2(x_1)$, set

$$C = \max [x \mid a < x < x_1, \qquad f_1(x) = f_2(x)],$$
$$D = \min [x \mid x_1 < x < b, \qquad f_1(x) = f_2(x)].$$

Then $f_1(C) = f_2(C), f_1(D) = f_2(D)$ and $C < D$. Now

$$\begin{aligned} f_1[F(C, D)] &= H[f_1(C), f_1(D), C, D] \\ &= H[f_2(C), f_2(D), C, D] = f_2[F(C, D)]. \end{aligned}$$

Here $C < F(C, D) < D$ by (2.3.1) and this shows the existence of an x_0 between C and D where $f_1(x_0) = f_2(x_0)$. This contradicts the definition of C and D and shows that $f_1(x) = f_2(x)$ in $[a, b]$.

Next we consider the interval $(b, B\rangle$. Let

$$E = \max [x \mid a \leqq x < B, \qquad f_1(x) = f_2(x)].$$

If $E < B$, there is an interval $(E, E + \epsilon)$ where $f_1 \neq f_2$. By (2.3.1) and the continuity of F, we can find an x_2 in this interval such that

$$a < F(x_2, a) < E.$$

We have then on the one hand

$$\begin{aligned} f_1[F(x_2, a)] &= H[f_1(x_2), f_1(a), x_2, a] \\ &\neq H[f_2(x_2), f_2(a), x_2, a] = f_2[F(x_2, a)], \end{aligned}$$

since H is strictly monotone in the first argument and $f_1(x_2) \neq f_2(x_2)$. On the other hand,

$$f_1[F(x_2, a)] = f_2[F(x_2, a)],$$

since f_1 and f_2 are equal in $[a, E]$. This contradiction shows that $E = B$, that is, $f_1 = f_2$ also in (b, B). In the same manner the remaining interval $\langle A, a)$ is handled.

It follows that the solution, if it exists and is continuous, is determined uniquely by the initial conditions (2.3.3).

Among the equations to which this argument applies is Jensen's equation

$$f\left(\frac{x+y}{2}\right) = \frac{1}{2}[f(x) + f(y)]. \tag{2.3.4}$$

Compare the inequality (1.6.2) for convex functions. Now (2.3.4) is obviously satisfied by a linear function

$$f(x) = \alpha x + \beta, \tag{2.3.5}$$

and such a function can be made to satisfy two conditions of type (2.3.2) which determine α and β uniquely. By Proposition 16 this is then the only continuous solution.

More sophisticated examples can be found in the theory of averages and in information theory (Aczél, *loc cit.*). It will mentioned later in Section 3.1 that each A-average corresponds to a continuous strictly monotone function F in terms of which the average is expressible. It is natural to ask if this function is uniquely determined. If we restrict ourselves to two variables, but generalize the question slightly, we are led to the functional equation

$$f[F(x, y)] = \frac{pf(x) + qf(y)}{p + q}, \tag{2.3.6}$$

where p and q are arbitrary positive numbers and

$$F(x, y) = \breve{g}\left(\frac{pg(x) + qg(y)}{p + q}\right). \tag{2.3.7}$$

Here g is a given continuous strictly monotone function and $\check{g}$ denotes its inverse. This equation is of type (2.3.2). It is not difficult to see that a solution is given by

$$f(x) = \alpha g(x) + \beta, \tag{2.3.8}$$

where α and β are arbitrary constants. Here again a condition of type (2.3.3) determines the constants uniquely and Proposition 16 shows that this is the only continuous solution taking on the prescribed initial values. It is clear that $f(x)$ is also strictly monotone so we have

$$\check{f}\left(\frac{pf(x) + qf(y)}{p+q}\right) = \check{g}\left(\frac{pg(x) + qg(y)}{p+q}\right). \tag{2.3.9}$$

A similar equation occurs in information theory, namely

$$f\left[\check{g}\left(\frac{xg(x) + yg(y)}{x+y}\right)\right] = \frac{xf(x) + yf(y)}{x+y}. \tag{2.3.10}$$

This is also of type (2.3.2) and here again the solution is of type (2.3.8).

Condition (2.3.1) is obviously rather restrictive and it is not satisfied by addition theorems. I am indebted to Professor Aczél for the information that he, in collaboration with M. Hosszú, has succeeded in replacing this condition by less restrictive assumptions. Thus, for the important case

$$f[F(x, y)] = H[f(x), f(y)], \tag{2.3.11}$$

there can be at most one continuous solution satisfying (2.3.3), if F is continuous and $F(x, y)$ and $H(u, v)$ are both strictly increasing (strictly decreasing) with respect to both variables involved.

2.4. *Miscellanea.* In this section we discuss briefly some questions involving functional equations which were discussed at the two Oberwolfach conferences. The first two questions deal with Abel's equation and are based on communications by Helmuth Kneser and A. Ostrowski, respectively.

The functional equation

$$F[f(x)] = F(x) + 1, \tag{2.4.1}$$

where f is a given function was studied by Abel [1], posthumous paper. It has been the object of much research. W. Schöbe [73] has given a solution of the special case

$$F(e^x) = F(x) + 1 \tag{2.4.2}$$

using a double iteration process. Set

$$L(x) = \log(1 + x), \qquad e(x) = e^x,$$

and define

$$L^n(x) = L[L^{n-1}(x)], \qquad e^n(x) = e[e^{n-1}(x)], \qquad n > 1.$$

Set

$$g(x) = \lim_{n\to\infty} \left\{ \frac{2}{L^n(x)} - n + \frac{1}{3}\log n \right\}, \tag{2.4.3}$$

$$h(x) = \lim_{n\to\infty} L^n[e^n(x)]. \tag{2.4.4}$$

The second limit exists for real values of x since the sequence involved is strictly increasing and bounded on compact sets. It is not known if $h(x)$ is analytic. The function g on the other hand is analytic. These functions satisfy the relations

$$g[L(x)] = g(x) + 1, \qquad h(x) = L[h(e(x))], \tag{2.4.5}$$

and

$$F(x) = -g[h(x)] \tag{2.4.6}$$

is a solution of (2.4.2). It is known that this equation has analytic solutions but it is not known if (2.4.6) defines such a solution.

Abel's equation (2.4.1) plays a role in the theory of convergence of infinite series, more precisely for the convergence criteria of V. P. Ermakov. In this first paper of 1871, Ermakov considered series of the form

$$\Sigma F(n), \tag{2.4.7}$$

where $F(x)$ is positive, continuous, and strictly decreasing. He showed that if a continuously differentiable function f could be

found such that $f(x) > x$, $f'(x) > 0$, and

$$\frac{F[f(x)]f'(x)}{F(x)} \leqq q < 1, \tag{2.4.8}$$

then the series converges, whereas if the quotient is $\geqq 1$ it diverges. In a later paper of 1883, he tried to eliminate the condition that $F(x)$ be monotone, though not with complete success. Correct conditions were proved by A. Ostrowski [65] in 1956. The main condition if that $f'(x)$ should be monotone, and if increasing it should attain or surpass unity. Following Ermakov he based the proof on the construction of a solution of Abel's equation with certain special properties.

At the 1962 conference W. Eichhorn (Würzburg) made an interesting application of the theory of hypercomplex numbers to a system of functional equations (see [90]). Let

$$F_i(s+t) = \sum_{j=1}^{n} \sum_{k=1}^{n} \gamma_i^{jk} F_j(s) F_k(t), \qquad i = 1, 2, \ldots, n. \tag{2.4.9}$$

He constructs an algebra A with basis elements $e^1, e^2, \ldots, e^n$ and the multiplication table

$$e^j e^k = \sum_{i=1}^{n} \gamma_i^{jk} e^i, \qquad j, k = 1, 2, \ldots, n. \tag{2.4.10}$$

Depending upon the "structure constants" γ_i^{jk}, A may be neither associative nor commutative. Multiplying the ith equation under (2.4.9) by e^i and summing for i, we get

$$F(s+t) = F(s)F(t), \qquad F = (F_1, F_2, \ldots, F_n). \tag{2.4.11}$$

If F is continuous and differentiable for all finite s, there exists an idempotent $b \in A$, $b^2 = b$, and an $a \in A$ such that $ba^k = a^k b = a^k$ for all natural numbers k and

$$F(s) = b \exp(as). \tag{2.4.12}$$

At the 1962 conference O. Taussky-Todd (Pasadena) raised the question of the solution of Cauchy's equation for quaternions or for Cayley numbers. At the 1963 conference some results of I. Makai (Debrecen) were announced which have a bearing on this

problem. If lower-case letters denote scalars and capitals quaternions, then the two equations

$$f(XY) = f(X)f(Y), \qquad F(xy) = F(x)F(y) \tag{2.4.13}$$

have as general solutions

$$f(X) = m(|X|), \qquad F(x) = m(x)[\cos \omega(x) + J \sin \omega(x)], \tag{2.4.14}$$

where

$$m(xy) = m(x)m(y), \qquad \omega(xy) \equiv \omega(x) + \omega(y) \pmod{2\pi},$$

and J is any fixed, unit quaternion with scalar part 0.

This finishes our survey of current events in the theory of functional equations.

MEAN VALUES

3.0. *Orientation.* Since Euclid proved that the geometric mean of two positive quantities cannot exceed their arithmetic mean, we can claim that considerations of mean values go back to antiquity. We have to wait until Cauchy and Gauss, however, for more sophisticated results, and the theory started to flourish only after L. Fejér [25] had shown in 1900 the importance of mean values for the theory of Fourier series. The study of special means, initiated by G. Frobenius [31] in 1880, O. Hölder [43] in 1882, and E. Cesàro [14] in 1890, ultimately led to a profusion of averaging processes applicable to the various needs of analysis.

Postulates for arithmetic means were given by R. Schimmack [87] in 1909 and these and other postulates were examined for logical independence by E. V. Huntington [85] in 1927 (references kindly supplied by J. B. Diaz and F. T. Metcalf). Complete determination of the mean values satisfying a given set of postulates was given in 1930 by A. N. Kolmogoroff [52] and M. Nagumo [62] and in 1931 by B. de Finetti [30]. This class of mean values will be discussed in the following and other applications will be found later in this report. It is by no means the only class worthy of study, but its general usefulness is unsurpassed.

There are three sections: *The Postulates. Consequences of the Postulates,* and *Remarks on Limitation.*

3.1 *The Postulates.* We consider an average subject to the following conditions.

(i) *For each n and each set of n positive numbers $x_1, x_2, \ldots, x_n$ there is a positive number $A(x_1, x_2, \ldots, x_n)$ called the A-average of these numbers.*

(ii) *$A(x_1, x_2, \ldots, x_n)$ is a continuous symmetric function of its arguments, strictly increasing in each of them.*

(iii) $A(x, x, \ldots, x) = x$.

(iv) *If $y = A(x_1, x_2, \ldots, x_k)$, then*

$$A(x_1, x_2, \ldots, x_k, x_{k+1}, \ldots, x_n) = A(y, y, \ldots, y, x_{k+1}, \ldots, x_n),$$

where y is repeated k times.

Here condition (iv) can be regarded as a class of functional equations satisfied by A. For $n = 3$, $k = 2$ this equation takes the form

$$A(x_1, x_2, x_3) = A[A(x_1, x_2), A(x_1, x_2), x_3]. \tag{3.1.1}$$

These various conditions naturally do not determine A in any way uniquely, but we have the following:

Proposition 17. To each A-average corresponds a strictly monotone function F, defined and continuous for $0 < u < \infty$, such that

$$A(x_1, x_2, \ldots, x_n) = \check{F}\left(\frac{1}{n}\sum_{j=1}^{n} F(x_j)\right), \tag{3.1.2}$$

where $\check{F}$ is the inverse of F.

Cf. remarks in Section 2.3 concerning the uniqueness of the association $A \leftrightarrow F$.

Additional conditions imposed on A tend to restrict F still further. Thus, if F is to be homogeneous of degree 1,

(v) $A(tx_1, tx_2, \ldots, tx_n) = tA(x_1, x_2, \ldots, x_n)$,

then $F(u)$ is either

$$u^\alpha \text{ or } \log u$$

with α real. If A is translation covariant

(vi) $A(x_1 + a, x_2 + a, \ldots, x_n + a) = A(x_1, x_2, \ldots, x_n) + a$,

then $F(u)$ is either

$$e^{\alpha u}, \qquad \alpha \neq 0, \qquad \text{or } u.$$

3.2. *Consequences of the Postulates.* Without using Proposition 17 one can prove a number of important properties of the averages directly from the postulates. We have

Proposition 18. Unless all the x's are equal

$$\min x_j < A(x_1, x_2, \ldots, x_n) < \max x_j. \tag{3.2.1}$$

This follows from (ii) + (iii).

Proposition 19. The average of k sets $x_1, x_2, \ldots, x_n$ equals the average of one set:

$$A(x_1, \ldots, x_1, x_2, \ldots, x_2, \ldots, x_n, \ldots, x_n) = A(x_1, x_2, \ldots, x_n). \tag{3.2.2}$$

Proof by rearrangement and repeated use of (iv).

This gives a convenient method of extending and contracting averages. The same method leads to the *principle of repeated averages*:

Proposition 20. From the n numbers $x_1, x_2, \ldots, x_n$ select k and form their average. The average of all the averages obtainable in this manner equals $A(x_1, x_2, \ldots, x_n)$.

To prove this we enlarge the given set E so as to obtain $\binom{n-1}{k-1}$ copies of each element x_j. The enlarged set E^* can be separated into $\binom{n}{k}$ distinct subsets each forming a selection of k elements of E. Averaging over E or over E^* gives the same result. On the other hand, in the average over E^* we can replace the elements of a subset by their average repeated k times. Thus the average over E^* equals the average of the $\binom{n}{k}$ averages each repeated k times. By Proposition 19 this reduces to the average of the averages as asserted.

This leads to a generalization of Proposition 18. Let

$$y_{k,m},\ m = 1, 2, \ldots, \binom{n}{k},$$

denote the averages of $x_1, x_2, \ldots, x_n$ taken k at a time.

Proposition 21. For each $k < n$

$$\min_m y_{k,m} < A(x_1, x_2, \ldots, x_n) < \max_m y_{k,m}, \tag{3.2.3}$$

unless all the x's are equal.

The last remark follows from the fact that the y's are equal if and only if the x's are equal.

This means that the averaging process is *oscillation reducing* in the following sense. Let S be an infinite bounded set of positive numbers. Select n distinct numbers from this set and form their A-average. The set of all such averages involving n numbers form a set S_n. For each n set

$$a_n = \inf S_n, \qquad A_n = \sup S_n.$$

Proposition 22. For all $n > 1$

$$a_{n-1} \leqq a_n \leqq A_n \leqq A_{n-1}, \tag{3.2.4}$$

so that

$$A_n - a_n \leqq A_{n-1} - a_{n-1}. \tag{3.2.5}$$

This follows from Proposition 21.

3.3. *Remarks on Limitation.* One of the most important applications of averaging processes is to the summation of infinite series or, equivalently, to the limitation of infinite sequences. Here the question of preservation of limits is basic. *The A-averages are limit preserving.* This follows from

Proposition 23. The A-average of k numbers a and n numbers b converges to b if $n \to \infty$ in such a way that $k/n \to 0$.

This is not so easy to derive directly from the postulates but it follows readily from Proposition 17. For if the average in question is $A_{k,n}$, then

$$F(A_{k,n}) = \frac{k}{k+n} F(a) + \frac{n}{k+n} F(b) \to F(b),$$

and since F is continuous and strictly monotone $A_{k,n} \to b$ as asserted.

Proposition 24. If $\lim x_n = x_0$, *then*

$$\lim A(x_1, x_2, \ldots, x_n) = x_0. \tag{3.3.1}$$

Suppose that $0 < \alpha < x_j < \beta$ for all j and that for a given $\epsilon > 0$ we have $0 < x_0 - \epsilon < x_j < x_0 + \epsilon$ for $j > k$. Then

$$A(\alpha, \ldots, \alpha, x_0 - \epsilon, \ldots, x_0 - \epsilon) < A(x_1, \ldots, x_k, x_{k+1}, \ldots, x_n) < A(\beta, \ldots, \beta, x_0 + \epsilon, \ldots, x_0 + \epsilon).$$

The first member converges to $x_0 - \epsilon$, the third to $x_0 + \epsilon$ whence the conclusion follows since the extension to the case $x_0 = 0$ is obvious.

We can obtain greater power and flexibility of averaging processes by composition. This was first observed by Otto Hölder [43] in 1882 for the case of the arithmetic means, $F(u) = u$. More generally, suppose that A_1 and A_2 are two A-averages, distinct or not, and form

$$[A_1 \cdot A_2](x_1, x_2, \ldots, x_n) \equiv A_1[A_2(x_1), A_2(x_1, x_2), \ldots, A_2(x_1, x_2, \ldots, x_n)]. \tag{3.3.2}$$

We have then the consistency theorem:

Proposition 25. If $\{x_n\}$ *is a sequence such that*

$$\lim A_2(x_1, x_2, \ldots, x_n) = x_0, \tag{3.3.3}$$

then also

$$\lim [A_1 \cdot A_2](x_1, x_2, \ldots, x_n) = x_0. \tag{3.3.4}$$

This follows from Proposition 24.

Naturally we have also that

$$A_1\text{-lim } x_n = y_0 \quad \text{implies } A_2A_1\text{-lim } x_n = y_0.$$

Since in general $A_1A_2 \neq A_2A_1$, we can not conclude that $y_0 = x_0$, that is, A_1 and A_2 need not be consistent.

We can clearly resort to compositions involving three or more "factors." Besides the Hölder means, not much use has been made of this device, one reason being the restriction to positive numbers which is undesirable for many applications.

Let us return to the oscillation-reducing property noticed above. This can be sharpened as follows. Let $\{x_n\}$ be a given sequence of positive numbers, $\{A(x)\}$ the sequence of successive A-averages. If

$$\begin{aligned} \xi_1 &= \liminf x_n, & \xi_2 &= \limsup x_n \\ \eta_1 &= \liminf A(x), & \eta_2 &= \limsup A(x), \end{aligned} \tag{3.3.5}$$

then using Propositions 21 and 23 we obtain

Proposition 26. We have

$$\eta_2 - \eta_1 \leqq \xi_2 - \xi_1. \tag{3.3.6}$$

This classical result has been widely extended to more general limiting processes and to complex-valued sequences and functions by K. Knopp [50]. His *Kernsatz* expresses that the convex hull of the cluster points of the transformed sequence is contained in the convex hull of the cluster points of the original sequence. Formula (3.3.6) is clearly a special case.

Knopp has also devoted much attention to the question of inequalities between different averages. See [51] and the literature quoted there. For the older literature see Hardy, Littlewood, and Pólya [34]. A typical result of Knopp's is

Proposition 27. Let F be twice differentiable and $F''(u) > 0$ in $(0, \infty)$. Let A be the average defined by $F(u)$ and M_1 the arithmetic mean defined by u. Let $x_1, x_2, \ldots, x_n$ lie in (a, b), $0 < a < b < \infty$. Then

$$0 \leqq F(A) - F(M_1) \leqq \lambda F(a) + (1 - \lambda)F(b) - F(\lambda a + (1 - \lambda)b), \tag{3.3.7}$$

where

$$\lambda = \frac{1}{b - a}\left\{b - g\left(\frac{F(b) - F(a)}{b - a}\right)\right. \text{ and } F'[g(u)] = u.$$

This does not include the inequality of P. Schweitzer [88] and L. V. Kantorovič [47]. If M_1 is the arithmetic, M_{-1} the harmonic mean of n numbers in $[a, b]$, then

$$\frac{M_1}{M_{-1}} \leqq \frac{(a+b)^2}{4ab} \tag{3.3.8}$$

For some recent results along these lines see Cargo and Shisha [9]. See also Diaz and Metcalf [84] to whom I am indebted for a number of references.

TRANSFINITE DIAMETERS

4.0. *Orientation.* In 1923 M. Fekete [26] investigated the distribution of the roots of a certain class of algebraic equations. He found that if the roots are restricted to a bounded closed set E in the complex plane, then the asymptotic behavior of the corresponding discriminants leads to a significant set function which he called the *transfinite diameter* of E. More precisely, he introduced

$$\delta_0(E) = \lim_{n\to\infty} \max \Big\{ \prod_{j<k} |z_j - z_k|^{2/(n(n-1))} \Big| \; z_1, z_2, \ldots, z_n \in E \Big\}. \tag{4.0.1}$$

The expression to be maximized is the geometric mean of the distances between the points $z_1, z_2, \ldots, z_n$ in E. This is the A-average corresponding to $F(u) = \log u$ in our previous notation. Fekete showed in a number of papers [27] that this concept is closely related to the Čebyšev constant of E, the equilibrium potential, and the exterior conformal mapping radius. See below.

In 1931 Pólya and Szegö [69] extended the investigation to R^3. They wanted to keep the connection with potential theory and found that this requires replacing the geometric mean by the harmonic mean based on the function $F(u) = u^{-1}$. They also considered general power means, $F(u) = u^p$.

A number of further generalizations were introduced by F. Leja in a series of papers starting in 1934. See in particular [54, 56] and the bibliography in [57]. Leja considered the connections with

conformal mapping, various associated sequences of polynomials, functions of two complex variables, and general metric spaces.

In the complex plane other transfinite diameters have been defined by Otto Frostman [32] and by M. Tsuji [78] using spherical or hyperbolical metrics.

There are three sections: *The Transfinite A-diameter*, *The Čebyšev Constant*, and *Special Cases.*

4.1. *The Transfinite A-diameter.* The following is based on Hille [38, 40] with additions and amendments.

Let X be a complete metric space, E a bounded set in X. Let A be an averaging process satisfying postulates (i)–(iv). Let $n > 1$ and take n distinct points $P_1, P_2, \ldots, P_n$ of E. Note the distances

$$\delta_{jk} = d(P_j, P_k), \qquad 1 \leqq j < k \leqq n.$$

Here $\delta_{jk} \leqq \delta(E)$, the topological diameter of E. Form the A-average of these distances, $A(\delta_{jk})$, which is also $\leqq \delta(E)$. Next we determine

$$\delta_n(E) = \sup A(\delta_{jk})$$

when the n points range over E. This leads to a bounded sequence $\{\delta_n(E)\}$ of positive numbers. This sequence is nonincreasing for any admissible A.

To show this we consider

$$A[d(Q_j, Q_k)], \qquad 1 \leqq j < k \leqq n + 1,$$

where the points Q_j are chosen so that the average exceed) $\delta_{n+1}(E) - \epsilon$ for a given $\epsilon > 0$. This is an average of $\frac{1}{2}n(n + 1$s numbers. Let us repeat each distance $n - 1$ times, thus obtaining a set of $\frac{1}{2}(n - 1)n(n + 1)$ numbers having the same average. We arrange these numbers into $n + 1$ groups of $\frac{1}{2}n(n - 1)$ numbers each such that the distances which do not involve the point Q_p are placed in the pth group. In forming the average, we can replace each element in the pth group by the average η_p of the elements in the group. This gives

$$A(\eta_1, \ldots, \eta_1, \eta_2, \ldots, \eta_2, \ldots, \eta_{n+1}, \ldots, \eta_{n+1}) = A(\eta_1, \eta_2, \ldots, \eta_{n+1}).$$

Hence

$$\delta_{n+1}(E) - \epsilon < A(\eta_1, \eta_2, \ldots, \eta_{n+1}) \leqq \max \eta_j \leqq \delta_n(E)$$

so that

$$\delta_{n+1}(E) \leqq \delta_n(E) \tag{4.1.1}$$

and

$$\lim_{n \to \infty} \delta_n(E) \equiv A\text{-}\delta_0(E) \tag{4.1.2}$$

exists. This is by definition *the transfinite A-diameter of E.*

This set function has important properties of monotony and continuity. The following is obvious:

Proposition 28. If $E_1 \subset E_2$, *then*

$$\delta_0(E_1) \leqq \delta_0(E_2). \tag{4.1.3}$$

This is the strongest statement that we can make. In fact, it often happens that E and its boundary ∂E have the same A-diameter.

Proposition 29. Let E_ϵ *be the set of points of* X *having a distance from* E *not exceeding* ϵ. *Then*

$$\lim_{\epsilon \downarrow 0} \delta_0(E_\epsilon) = \delta_0(E). \tag{4.1.4}$$

To show this, choose an arbitrary integer n, let $\eta > 0$ be given and find points $Q_1, \ldots, Q_n$ in E_ϵ whose average distance $A(Q) \geqq \delta_n(E_\epsilon) - \eta$. For each point Q_j we can find a point $P_j \in E$ such that $0 < d(P_j, Q_j) \leqq \epsilon$, $P_j \neq P_k$ if $j \neq k$. Then

$$d(Q_j, Q_k) \leqq d(P_j, P_k) + 2\epsilon.$$

Thus

$$\delta_n(E_\epsilon) - \eta \leqq A[d(P_j, P_k) + 2\epsilon].$$

Now the function $A(x)$ is continuous in each of its $\frac{1}{2}n(n-1)$ arguments at the point $\{\ldots, d(P_j, P_k), \ldots\}$. This means that a $\Delta(\epsilon)$ can be found which goes to 0 with ϵ such that

$$A[d(P_j, P_k) + 2\epsilon] < A[d(P_j, P_k)] + \Delta(\epsilon) \leqq \delta_n(E) + \Delta(\epsilon).$$

Hence

$$\delta_n(E_\epsilon) \leqq \delta_n(E) + \Delta(\epsilon) + \eta.$$

Here the left member is $\geqq \delta_0(E_\epsilon)$, so we have

$$\delta_0(E_\epsilon) \leqq \delta_n(E) + \Delta(\epsilon) + \eta.$$

The left member is a nondecreasing function of ϵ so it tends to a limit as $\epsilon \to 0$ and $\Delta(\epsilon) \to 0$. Hence

$$\lim_{\epsilon \downarrow 0} \delta_0(E_\epsilon) \leqq \delta_n(E) + \eta.$$

for every n. Here we let $n \to \infty$ and obtain

$$\lim \delta_0(E_\epsilon) \leqq \delta_0(E) + \eta$$

for every $\eta > 0$ and hence also for $\eta = 0$. Since the limit is at least $\delta_0(E)$, formula (4.1.4) follows.

4.2. *The Čebyšev Constant.* Let E be a bounded closed set in the complex plane and let $P_n(z)$ be any polynomial of degree n and leading term z^n. The absolute value of $P_n(z)$ has a maximum in E which is attained. In other words, the geometric mean of the distances from the variable point z to the zeros of $P_n(z)$ has a maximum in E. The set of all such maxima for P_n, ranging over all admissible polynomials, is bounded below and the infimum is reached. There is a unique polynomial $T_n(z)$, called the *nth Čebyšev polynomial for* E such that

$$(4.2.1) \qquad \max_{z \in E} |T_n(z)|^{1/n} = \min_{P_n} \max_{z \in E} |P_n(z)|^{1/n} \equiv M_n(E).$$

The classical *Čebyšev constant of* E is then by definition

$$(4.2.2) \qquad \chi(E) = \lim_{n \to \infty} M_n(E).$$

The existence of this limit follows from the general argument given below. Fekete proved that

$$(4.2.3) \qquad \chi(E) = \delta_0(E).$$

This construction presupposes that the given set is in the complex plane and that the geometric mean is used to form the average. But we can carry over the procedure to any complete metric space, any bounded set E, and any A-average. The geometric interpretation makes sense in this new setting and we have merely to replace "min max" by "inf sup." The equality (4.2.3) is usually replaced by an inequality.

We proceed as follows. Take n points $P_1, \ldots, P_n$ in the space X and form the function

$$f(P) = A[d(P, P_1), d(P, P_2), \ldots, d(P, P_n)], \tag{4.2.4}$$

that is the average of the distance of the variable point P from the fixed points $P_1, \ldots, P_n$. This function has a supremum on the bounded set E. The set of all such suprema corresponding to n points $P_j \in X$ is bounded below and we denote its infimum by $M_n(E)$. Thus

$$M_n(E) = \inf \sup_{P \in E} f(P). \tag{4.2.5}$$

If E is compact, we can actually find a function $f_n(P)$ which assumes the value $M_n(E)$ so that "inf sup" becomes "min max." This function need not be uniquely determined.

The sequence $\{M_n(E)\}$ is convergent. This follows from the inequalities

$$M_{m+n} \leqq \max (M_m, M_n), \tag{4.2.6}$$

$$M_{n+1} \leqq A(M_n, \ldots, M_n, b), \tag{4.2.7}$$

where M_n occurs n times and b is any number $\geqq \delta(E)$.

Here the first inequality is of type (1.6.8) with $S = Z_+^1$. To verify it, note that we can find functions

$$f_1(P) = A[d(P, P_1), \ldots, d(P, P_m)],$$

$$f_2(P) = A[d(P, Q_1), \ldots, d(P, Q_n)]$$

whose suprema on E differ from $M_m(E)$ and $M_n(E)$, respectively, by less than a preassigned ϵ. We then form

$$A[d(P, P_1), \ldots, d(P, P_m), d(P, Q_1), \ldots, d(P, Q_n)]$$

whose supremum on E is at least $M_{m+n}(E)$. On the other hand, for all P this expression equals

$$A[f_1(P), \ldots, f_1(P), f_2(P), \ldots, f_2(P)] \leqq \max [f_1(P), f_2(P)]$$

and on E this is at most equal to the larger of the numbers $M_m + \epsilon$, $M_n + \epsilon$. This proves (4.2.6) and (4.2.7) is proved by a similar argument.

The first inequality shows that $M_n(E) \leqq M_1(E)$ for all n so

$\{M_n(E)\}$ is a bounded sequence. Set

$$\limsup M_n = \beta.$$

If there is a j and a number γ such that

$$M_j < \gamma, M_{j+1} < \gamma,$$

then (4.2.6) shows that $M_k < \gamma$ for all large k ($k > j^2$ will do). It follows that such a number γ must be $\geqq \beta$. Suppose now that there is a j and a γ such that $M_j < \gamma < \beta$. Then by (4.2.6) $M_k < \gamma$ for infinitely many values of k, for instance for $k = j2^m$. Then (4.2.7) shows that for such a k

$$M_{k+1} \leqq A(M_k, \ldots, M_k, b) < A(\gamma, \ldots, \gamma, b),$$

where M_k and γ are each repeated k times. By Proposition 23 the last member is arbitrarily close to γ for large values of m. Thus we again obtain two consecutive numbers M_k and M_{k+1}, both less than a fixed number $<\beta$ and we have again a contradiction. This shows that

$$M_n(E) \geqq \limsup M_k(E) \tag{4.2.8}$$

for all n and

$$\lim M_n(E) \equiv \chi(E) \tag{4.2.9}$$

exists.

We have

$$\delta_0(E) \geqq \chi(E) \tag{4.2.10}$$

and strict inequality may hold.

To show this, choose n points $P_1, \ldots, P_n$ in E such that for a given $\epsilon > 0$

$$A[d(P_j, P_k)] = \delta_n(E) - \epsilon.$$

Then choose a point $P_{n+1} \in E$ such that

$$A[d(P_{n+1}, P_1), \ldots, d(P_{n+1}, P_n)] \geqq M_n(E).$$

This is possible. Then form the average of all distances between the $n + 1$ points. This is at most equal to $\delta_{n+1}(E)$. On the other hand, the distances form two groups, those involving P_{n+1} and those that do not. The average for the first group is $\geqq M_n(E)$,

that for the second is $\delta_n(E) - \epsilon$. It follows that

$$\delta_{n+1}(E) \geqq \min [\delta_n(E) - \epsilon, M_n(E)].$$

This must hold for all ϵ so we have

$$\delta_{n+1}(E) \geqq \min [\delta_n(E), M_n(E)] = M_n(E)$$

for all n and (4.2.10) follows.

Examples for the inequality are found in Pólya and Szegö [69]. Thus, for instance, in R^3 with E taken as the unit ball and A being pth power means, $F(u) = u^p$ with $-1 < p < 2$, we have

$$(4.2.11) \quad A\text{-}\delta_0(E) = 2(1 + \tfrac{1}{2}p)^{-1/p} > A\text{-}\chi(E) = 1.$$

4.3. *Special Cases.* For the pth power means Pólya and Szegö listed transfinite diameters and Čebyšev constants for the following configurations: two points, a line segment, a circle, a circular disk, the surface of a sphere, and the solid sphere. They also indicated the nature of the equilibrium distribution on these objects for the corresponding potentials.

The problem of finding the A-diameter of the unit sphere in a Banach space leads to some interesting results [40]. Let X be a Banach space over complex numbers. Let E be a c-star with respect to the origin, that is, $x_0 \in E$ implies $\zeta x_0 \in E$ for any complex number ζ with $|\zeta| \leqq 1$. If ρ is the A-diameter of the unit disk in the complex plane and if $r = \sup \|x_0\|$, $x_0 \in E$, then

$$(4.3.1) \qquad A\text{-}\delta_0[E] \geqq \rho r.$$

This holds in particular if E is the unit sphere and here $r = 1$.

In any Banach space X with unit sphere U

$$(4.3.2) \qquad A\text{-}\delta_0[U] \leqq 2,$$

and equality holds in a number of cases, in particular for

$$X : C[a, b], L^1(a, b), L^\infty(a, b), l, m.$$

The Lebesgue spaces $L^p(0, 1)$, $1 < p < \infty$, present some unsolved problems. In such a space the elements of the orthonormal system [exp (kix)] are equidistant. This gives

$$(4.3.3) \quad A\text{-}\delta_0[U^{(p)}] \geqq 2\left[\frac{\Gamma(\frac{1}{2}p + \frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}p + 1)}\right]^{1/p}, \qquad 1 \leqq p \leqq \infty.$$

This can be improved to

$$A\text{-}\delta_0[U^{(p)}] \geqq 2^{1/p}, \qquad 1 \leqq p \leqq 2. \tag{4.3.4}$$

Consider the Euclidean n space, R^n, and its unit ball U^n. If a certain choice of n unit vectors gives an A-average equal to A-$\delta_n[U^n]$, then these vectors also lie in U^{n-1}. For in any case their endpoints determine a hyperplane the equation of which can be taken to be $x_n = \rho$ with $0 \leqq \rho < 1$. If the maximizing vectors are

$$v_k = (x_{1,k}, \ldots, x_{n-1,k}, \rho), \qquad k = 1, 2, \ldots, n,$$

and if $\rho > 0$, then

$$u_k = (1 - \rho^2)^{-1/2}(x_{1,k}, \ldots, x_{n-1,k}, 0)$$

are unit vectors and

$$\|u_j - u_k\| = (1 - \rho^2)^{-1/2}\|v_j - v_k\|,$$

so that

$$A[\|u_j - u_k\|] > A[\|v_j - v_k\|] = \delta_n(U^n)$$

which is absurd. Hence $\rho = 0$ and the maximizing vectors belong to U^{n-1}. This shows that

$$A\text{-}\delta_n[U^n] = A\text{-}\delta_n[U^{n-1}]. \tag{4.3.5}$$

POTENTIAL THEORY

5.0. *Orientation.* We shall discuss some modern aspects of general potential theory in R^m, $m \geqq 1$. Let $x = (x_1, x_2, \ldots, x_m)$ denote a point in R^m and write $\|x\|$ for the length of the vector x.

Classical potential theory is based on the use of certain kernels which satisfy Laplace's equation, $\log(1/r)$ for $m = 2$ and r^{2-m} for $m \neq 2$. The use of more general kernels goes back to Pólya and Szegö [69], M. Riesz [70], and Frostman [32] in the early 1930's. The relation between capacity and transfinite diameter goes back to Szegö [76] in 1924. For the general notion of capacity see C. Choquet [15] and L. Carleson [11]. The latter's paper on exceptional sets contains a bibliography of close to 600 items. See also E. Cartan [12], J. Deny [18], and K. Kunugui [53].

There are five sections: *Capacity, Capacitary and Equilibrium Potentials, Green's Function, Some Special Kernels,* and *A Problem.*

5.1. *Capacity.* Let $K(r)$ be a positive nonincreasing continuous function defined for $r > 0$ with $\lim_{r \downarrow 0} K(r) \leqq \infty$. If σ is a measure of compact support in R^m, the *K-potential* of σ is by definition

$$u(x \mid \sigma, K) = \int K(\|x - y\|)\, d\sigma(y), \tag{5.1.1}$$

and the corresponding *energy integral* is

$$I(\sigma, K) = \int \int K(\|x - y\|)\, d\sigma(y)\, d\sigma(x), \tag{5.1.2}$$

the integrals taken over the support of σ.

If E is a compact set in R^m and $\Gamma(E)$ is the set of all positive measures μ with total mass 1 concentrated on E, then

$$C_K(E) = \{\inf I(E, \mu, K)\}^{-1}, \mu \in \Gamma(E), \tag{5.1.3}$$

is the *K-capacity* of E. In particular, the cases $K(r) = \log(1/r)$ and $r^{-\alpha}$ give rise to *logarithmic capacity* and *α-capacity*, respectively.

For a given K, there are two possibilities: (i) $C_K(E) = 0$ or (ii) $0 < C_K(E) < \infty$. The problem of characterizing sets E with $C_K(E) = 0$ in terms of other properties has received much attention.

At this juncture let us recall the definition of *Hausdorff measure.* Let $h(t)$ be a continuous increasing function on $[0, 1)$ with $h(0) = 0$, $\int_0^\omega [h(t)/t]\, dt < \infty$. Let ρ be given and cover the set E by a finite or countably infinite set of spheres S_j where the radius of S_j is $r_j \leqq \rho$ and set

$$H(\rho) = \inf \Sigma\, h(r_j)$$

for all possible coverings of E. This is a nondecreasing function of ρ and

$$m_h(E) = \lim_{\rho \downarrow 0} H(\rho) \tag{5.1.4}$$

is the Hausdorff measure of E with respect to the measuring function h. It is an outer measure in the sense of Carathéodory and hence a countably additive set function on Borel sets. The case $h(t) = t^\alpha$ gives α-dimensional measure, $[\log(1/t)]^{-1}$ logarithmic measure (integrability condition not satisfied!).

It was proved by Frostman [32] that if a closed set E in the plane

has logarithmic capacity 0 then its Hausdorff measure is 0 with respect to each admissible measuring function h. He also showed that the logarithmic capacity is positive if some Hausdorff measure is positive, and this also extends to α-capacities provided the Hausdorff measure satisfies the stronger restriction that $h(t)t^{-1-\alpha} \in L(0, 1)$. In the other direction Carleson [10] gave an example of a closed set E of positive logarithmic capacity such that every admissible Hausdorff measure vanishes. He also gave a simple proof of a result of Erdös and Gillis [20] according to which a set of finite logarithmic measure always is of logarithmic capacity 0. This shows that these two ways of measuring sets are fundamentally different.

An interesting limit theorem has recently been proved by Hans Wallin [82]. Let E be a compact set in R^m and let $M[E]$ be its m-dimensional Lebesgue measure, $C_\alpha[E]$ its α-capacity, $\alpha < m$. Then

$$\lim_{\alpha \uparrow m} \frac{C_\alpha(E)}{m - \alpha} = kM(E), \tag{5.1.5}$$

where k is a function of m alone.

Evans [23] proved that if a compact set E in R^m has α-capacity 0, $0 \leqq \alpha < m$ and $\alpha = 0$ means logarithmic capacity, then there exists a positive measure μ on E such that the α-potential of μ is infinite everywhere on E. Wallin (*loc. cit.*) proved that E has α-capacity 0 if and only if every function f continuous in E coincides in E with a continuous α-potential of a measure with compact support. This potential is continuous in the whole space and belongs to C^∞ outside of E. This result holds also for more general kernels K. Further he shows that if f is continuous in E and $C_\alpha(E) = 0$, then f can be extended to a function whose partial derivatives of certain orders belong to $L^p(R^m \ominus E)$ where the orders depend upon α and p. See also [81] where the following result is proved. Let U^m be the open unit sphere in R^m, $m \geqq 2$, and E a closed subset of U^m. Then $C_{m-2}(E) = 0$ if and only if every f continuous on E can be extended to a function u, harmonic on U^m and continuous on its closure, which has a finite Dirichlet integral over U and coincides with f in E.

5.2. *Capacitary and Equilibrium Potentials.* For questions where zero capacity or positive capacity is the only property that matters,

formula (5.1.3) which goes back to M. Riesz [70] gives the most convenient definition of K-capacity. But for connections with other fields it is better to use the definition of Frostman [32]. To distinguish between the two varieties, that of Frostman will be denoted by $C_K^*(E)$ and called the star capacity. It is defined by

$$(5.2.1)\quad K[C_K^*(E)] = \inf I(E, \mu, K) \equiv V_K(E), \qquad \mu \in \Gamma(E).$$

If $K(0) = \infty$, the two definitions are consistent in the sense that $C_K(E) = 0$ if and only if $C_K^*(E) = 0$.

Suppose now that $C_K^*(E) > 0$. It is then possible to find a sequence of positive measures $\mu_n \in \Gamma(E)$ such that

(i) $$I(E, \mu_n, K) \to V_K(E).$$

(ii) There is a $\mu_0 \in \Gamma(E)$ such that

$$(5.2.2)\qquad \int_E f(x)\, d\mu_n(x) \to \int_E f(x)\, d\mu_0(x)$$

for every f continuous in E.

(iii) $$I(E, \mu_0, K) = V_K(E).$$

If (5.2.2) holds, $\{\mu_n\}$ is said to converge weakly to μ_0. If μ_0 is unique, it is called the *equilibrium distribution* on E and $u(x \mid \mu_0, K)$ is called the *equilibrium potential*. Uniqueness will occur if ∂E is sufficiently regular. In any case $u(x \mid \mu_0, K)$ is a *capacitary potential* and μ_0 is a *capacitary distribution*.

If μ_0 is unique it can be obtained in the following manner. K is strictly monotone for $r > 0$ and defines an admissible A-average A_K by formula (3.1.2) replacing F by K. Since E is compact we can find points $x_1, x_2, \ldots, x_n$ in E such that

$$(5.2.3)\qquad A_K(\|x_j - x_k\|) = \delta_n(E),$$

$$(5.2.4)\qquad \binom{n}{2} K[\delta_n(E)] = \sum_{1 \le j < k \le n} K(\|x_j - x_k\|).$$

At each of the points x_j we place a mass $1/n$. This defines a $\mu_n \in \Gamma(E)$ and

$$(5.2.5)\quad I(E, \mu_n, K) = K[\delta_n(E)] \to K[\delta_0(E)], \qquad n \to \infty.$$

Hence

$$A_K\text{-}\delta_0(E) \leqq C_K^*(E), \tag{5.2.6}$$

and it may be shown that equality holds. Szegö [76] proved equality for the logarithmic case, Frostman [32] for $K = r^{-\alpha}$.

If $C_K^*(E) = 0$, it is no longer necessarily true that the sequence $\{\mu_n\}$ converges to a limit. For the case $K = r^{-1}$ Terasaka [77] contructed a closed countable set E such that the sequence

$$u_n(x) = \int_E K(\|x - y\|)\, d\mu_n(y)$$

does not converge for every x in $C(E)$, that is, μ_n does not converge weakly. Wallin [81] has observed that such a construction can be carried through for every admissible kernel K.

If ν is a capacitary distribution for E, then

$$u(x) = u(x \mid \nu, K)$$

has the following properties:

(i) $u(x) \geqq V_K(E)$ for all $x \in E$ except perhaps for a set of K-capacity 0.

(ii) $u(x) \leqq V_K(E)$ everywhere on the support of ν.

(iii) $u(x) \leqq AV_K(E)$ for all x, where the constant A depends only on m.

Here the first two properties follow from the variation principle of Gauss, modified by Frostman, whereas the third is due to Ugaheri [79].

Continuity properties of $C_K(E)$ with respect to K have been examined by Wallin [81]. If $K_n \to K$ and $C_{K_n}(E) > 0$ while $C_K(E) = 0$, and if μ_n is a capacitary distribution for E and K_n, does $\{\mu_n\}$ converge weakly to a limit? This turns out to be the case if E has positive m-dimensional Lebesgue measure and $K_n(r)r^{m-1} \in L(0, 1)$ while $K(r)r^{m-1}$ does not. Various plausible conjectures for such problems are disproved by counterexamples.

5.3. *Green's Function.* Let E be a compact set of positive K-capacity in R^m and suppose that μ_0 is the equilibrium distribution of E. Let z be exterior to E. The *function of Green relative to the set E, the point z, and the kernel K* is of the form

$$G(x, z; E) = K(\|z - x\|) - \int_E K(\|y - x\|)\, d_y\nu(y, z). \tag{5.3.1}$$

Here the measure $\nu(y, z)$ is to be chosen so that

$$G(x, z; E) \geqq 0$$

everywhere, with equality in E excepting at most a set of points x of capacity zero.

Since

$$\int_E \int_E K(\|s - t\|)\, d_s\nu(s, x)\, d_t\nu(t, z)$$

is a symmetric function of x and z, one obtains that so is G:

$$G(x, z; E) = G(z, x; E). \tag{5.3.2}$$

The existence of such a function is of basic importance for potential theory. For if such a function exists,

$$\int_E G(x, z)\, d\mu(x) = 0, \qquad \mu \in \Gamma(E),$$

implies that

$$u(z) = \int_E u(y)\, d_y\nu(y, z) \tag{5.3.3}$$

for any potential function

$$u(x) = \int_E K(\|x - y\|)\, d\mu(y). \tag{5.3.4}$$

The existence of $G(x, z; E)$ has been proved for special kernels K under rather general assumptions on E. For the case of the Newtonian potential, $K(r) = r^{-1}$, there is a proof in Frostman [32], for the more general kernel $r^{-\alpha}$ in M. Riesz [70]. We shall devote the remainder of this section to the logarithmic case.

We assume that the compact set E in the complex plane has a (unique) equilibrium distribution μ_0 with respect to $\log(1/r)$ and that $V(E)$ is the minimum value of the energy integral. Then

$$G(z, \infty; E) \equiv V(E) - u(z \mid \mu_0, E) \tag{5.3.5}$$

is by definition Green's function for μ_0 and $C(E)$, the complement of E. It is harmonic in $C(E)$, tends to zero as z tends to any point on ∂E excepting a set of capacity zero, and it becomes infinite as $\log|z|$ when $|z| \to \infty$.

For an arbitrary point $z_0 \in C(E)$ we proceed as follows. The

transformation

$$w = \frac{1}{z - z_0} \tag{5.3.6}$$

maps E into a compact set F in the w-plane. It takes functions harmonic in $C(E) \ominus \{\infty\}$ into functions harmonic in $C(F) \ominus \{z_0\}$. Let μ_1 be the equilibrium distribution on F and let $V(F)$ be the minimum of the energy integral. Then

$$G(w, \infty; F) = V(F) - u(w \mid \mu_1, F)$$

and we set

$$G(z, z_0; E) = G(w, \infty; F). \tag{5.3.7}$$

This function evidently has the desired properties and since Green's function is unique when it exists, it follows that this is the correct definition.

Let E be a bounded continuum whose complement $C(E)$ is simply connected. Szegö had observed that the logarithmic capacity of E coincides with the (geometric) transfinite diameter and Fekete [27] could prove that the common value agrees with that of the exterior mapping radius of E.

Let $F_n(z; E)$ be the nth Fekete polynomial of E, that is,

$$F_n(z; E) = F_n(z) = \prod_{j=1}^{n} (z - z_{j,n}) \tag{5.3.8}$$

and

$$\prod_{j<k} |z_{j,n} - z_{k,n}| = \{\delta_n(E)\}^{\binom{n}{2}}.$$

Set

$$f_n(z) = [F_n(z)]^{1/n} \tag{5.3.9}$$

with the nth root so chosen that

$$\lim_{z \to \infty} z^{-1} f_n(z) = 1.$$

A basic property of the Fekete polynomials is that

$$\lim_{n \to \infty} \max_{z \in E} |f_n(z)| = \delta_0(E), \tag{5.3.10}$$

the (geometric) transfinite diameter of E.

The functions $\{f_n(z)\}$ are holomorphic in $C(E)$ and

$$\lim_{n\to\infty} f_n(z) \equiv f(z) \tag{5.3.11}$$

exists uniformly on compact subsets of $C(E)$. The limit function maps $C(E)$ conformally on

$$|w| > \delta_0(E),$$

and since $f(z)$ has an expansion of the form

$$f(z) = z + \sum_{k=1}^{\infty} A_k z^{1-k} \tag{5.3.12}$$

for large values of $|z|$, it follows that $\delta_0(E)$ equals the exterior conformal mapping radius of E as asserted.

On the other hand, if μ_n is the distribution of mass on E where each point $z_{j,n}$ gets the mass $1/n$, then

$$-\log|f_n(z)| = \int_E \log\frac{1}{|z-t|}\, d\mu_n(t).$$

As $n \to \infty$ the left member converges to $-\log|f(z)|$ and μ_n converges weakly to the equilibrium distribution on E so that

$$-\log|f(z)| = \int_E \log\frac{1}{|z-t|}\, d\mu_0(t) = u(z \mid \mu_0, E), \tag{5.3.13}$$

the equilibrium potential on E. Here all three members are harmonic in $C(E)$; the third member approaches $V(E)$ on ∂E, excepting at most a set of capacity 0, whereas the first member approaches $-\log\delta_0(E)$ almost everywhere on ∂E. It follows that

$$-\log\delta_0(E) = V(E) = -\log C^*(E)$$

or

$$\delta_0(E) = C^*(E) \tag{5.3.14}$$

as asserted above. We see also that

$$G(z, \infty; E) = \log|f(z)| - \log C^*(E). \tag{5.3.15}$$

In view of the preceding construction Green's function, $G(z, z_0; E)$ as well as $G(z, \infty; E)$, can be obtained by the same averaging process

as gives the transfinite diameter. It is not clear if a similar result holds in R^3 for the Newtonian potential.

The coefficients A_k of (5.3.12) are expressible in terms of the moments of the equilibrium distribution

$$M_n = \int_E t^n \, d\mu_0(t). \tag{5.3.16}$$

We have

$$\frac{f'(z)}{f(z)} = \sum_{k=0}^{\infty} M_k z^{-k-1}, \tag{5.3.17}$$

and

$$A_k = \sum \frac{(-1)^{p_1+p_2+\cdots+p_k}}{p_1!p_2!\ \ldots\ p_k!}\left(\frac{M_1}{1}\right)^{p_1}\left(\frac{M_2}{2}\right)^{p_2}\cdots\left(\frac{M_k}{k}\right)^{p_k}, \tag{5.3.18}$$

where the summation extends over those non-negative integers p_j such that

$$p_1 + 2p_2 + \cdots + kp_k = k.$$

The expression of A_k in terms of M_j coincides with the Newton-Waring formula for the kth symmetric function of n variables, $n > k$, in terms of the corresponding power sums.

The argument given above goes back to F. Leja [55]. See also Hille [37, II, p. 339 et seq.]. The Fekete polynomials $F_n(z; E)$ can be replaced by the corresponding Čebyšev polynomials with the same result. In fact, it follows from G. Faber [24] that the mapping function is the limit of nth roots of other polynomials associated with the compact set E.

5.4 *Some Special Kernels.* We return again briefly to m-dimensional space R^m. Here the power kernels have an important composition property which becomes more striking if we modify the definitions. We set

$$R_\alpha(x) = \frac{\Gamma(\frac{1}{2}m - \frac{1}{2}\alpha)}{\pi^{\frac{1}{2}m}\Gamma(\frac{1}{2}\alpha)}\|x\|^{\alpha-m}, \qquad 0 < \alpha < m. \tag{5.4.1}$$

This is the original definition of M. Riesz [70] who was led to it in an effort to generalize the fractional, Riemann-Liouville integral in one

dimension. The composition property is based on the semigroup property of the kernel with respect to the parameter

$$R_{\alpha+\beta}(x) = R_\alpha * R_\beta(x), \qquad 0 < \alpha,\ 0 < \beta,\ \alpha + \beta < m, \tag{5.4.2}$$

or written out in full

$$R_{\alpha+\beta}(x) = \int_{R^m} R_\alpha(s) R_\beta(x - s)\, ds. \tag{5.4.3}$$

This kernel gives an α-potential

$$u(x \mid \mu, \alpha) = \int_E R_\alpha(x - y)\, d\mu(y) \tag{5.4.4}$$

with the composition property

$$u(x \mid \mu, \alpha + \beta) = u[u(x \mid \mu, \alpha) \mid \mu, \beta]. \tag{5.4.5}$$

If Δ denotes the Laplacian in R^m we have also

$$\Delta u(x \mid \mu, \alpha + 2) = -u(x \mid \mu, \alpha). \tag{5.4.6}$$

The Riesz potentials suffer from one defect, namely the limitation imposed on the parameter α. Riesz has indicated how this handicap can be overcome by analytic continuation with respect to α provided the mass distribution is sufficiently regular.

Another device has been used successfully by N. Aronszajn and K. T. Smith [6] in a recent study of what they call *Bessel potentials*. These are based on the kernel

$$G_\alpha(x) = \frac{2^{\frac{1}{2}(2-m-\alpha)}}{\pi^{m/2}\Gamma(\frac{1}{2}\alpha)} K_{\frac{1}{2}(m-\alpha)}(\|x\|)\|x\|^{\frac{1}{2}(\alpha-m)}. \tag{5.4.7}$$

Here

$$K_\nu(z) = \frac{\pi}{2}\frac{I_{-\nu}(z) - I_\nu(z)}{\sin \nu\pi}, \qquad I_\nu(z) = \sum_{k=0}^{\infty} \frac{(\frac{1}{2}z)^{\nu+2k}}{k!\Gamma(\nu + k + 1)},$$

so that $K_\nu(z)$ is a modified Bessel function of the third kind. At the origin $G_\alpha(x)$ behaves as $R_\alpha(x)$ for $\alpha < m$ but is bounded for $\alpha > m$ and it goes exponentially to 0 as $\|x\| \to \infty$. This implies that the Bessel potential can be defined for all $\alpha > 0$. We have also

$$G_{\alpha+\beta}(x) = G_\alpha * G_\beta(x) = \int_{R^m} G_\alpha(x - y) G_\beta(y)\, dy \tag{5.4.8}$$

valid for all positive α, β.

5.5. *A Problem.* In conclusion let us make a remark concerning a special polynomial series. Let $P_n(z)$ denote either $F_n(z; E)$ or $T_n(z; E)$ where E is a compact set in the complex plane, Int (E) is connected and $\neq \emptyset$, and C(E) is simply connected. Let

$$\limsup |a_n|^{1/n} = \frac{1}{\delta_0(E)}, \tag{5.5.1}$$

then the series

$$a_0 + \sum_{n=1}^{\infty} a_n P_n(z) \tag{5.5.2}$$

is absolutely convergent in Int (E) and diverges everywhere in C(E). Its sum is holomorphic in Int (E).

Is every function which is holomorphic in Int (E) representable by a series (5.5.2) *and, if so, is the representation unique?*

REFERENCES

1. N. H. Abel, Determination d'une fonction au moyen d'une équation qui ne contient qu'une seule variable, *Œuvres*, II (1881) 36–39.
2. J. Aczél, *Vorlesungen über Funktionalgleichungen und ihre Anwendungen*, Basel (1961) Birkhäuser.
3. J. Aczél, Ein Eindeutigkeitssatz in der Theorie der Funktionalgleichungen und einige ihrer Anwendungen, *Acta Math. Acad. Sci. Hungar.*, **15** (1964).
4. L. V. Ahlfors, On Phragmén-Lindelöf's principle, *Trans. Amer. Math. Soc.*, **41** (1937) 1–8.
5. A. A. Albert, *Modern higher algebra*, Chicago (1937) Univ. Chicago Press.
6. N. Aronszajn and K. T. Smith, Theory of Bessel potentials, I., *Ann. Inst. Fourier, Grenoble*, **11** (1961) 385–475.
7. E. Beckenbach and R. Bellman, *Inequalities*, Berlin (1961) Springer [*Ergebnisse d. Math.*, N.F., **30**].
8. N. Bourbaki, Fonctions d'une variable réele, Ch. IV. *Act. Sci. Ind.*, 1132, Hermann, Paris, 1951.
9. G. T. Cargo and O. Shisha, Bounds on ratios of means, *J. Res. Nat. Bur. Standards*, B, **66 B**:4 (1962) 169–170.
10. L. Carleson, On the connection between Hausdorff measures and capacity, *Arkiv f. Mat.*, **3**:36 (1957) 403–406.
11. L. Carleson, *Selected problems on exceptional sets*, Uppsala (1961) 81 pp. (mimeo.)
12. H. Cartan, Sur les fondements de la théorie du potential, *Bull. Soc. Math. France*, **69** (1941) 71–96.
13. A. L. Cauchy, *Cours d'Analyse de l'École Royale Polytechnique*, 1. *Analyse algébrique*, Paris (1821) [*Œuvres* (2) **3** (1897) 98–113, 220].

14. E. Cesáro, Sur la multiplication des séries, *Bull. Sci. Math.*, (2) **14** (1890) 114–120.
15. G. Choquet, Theory of capacities, *Ann. Inst. Fourier, Grenoble*, **5** (1953–1954) 131–295.
16. G. Choquet and J. Deny, Sur l'équation de convolution $\mu = \mu * \sigma$. *C.R. Acad. Sci. Paris*, **250** (1960) 799–801.
17. G. Darboux, Sur la composition des forces en statique, *Bull. Sci. Math.*, (1) **9** (1875) 281–288.
18. J. Deny, Les potentiels d'énergie finie, *Acta Math.*, **82** (1950) 107–183.
19. N. Dunford and E. Hille, The differentiability and uniqueness of continuous solutions of addition formulas, *Bull. Amer. Math. Soc.*, **50** (1944) 799–805.
20. P. Erdös and J. Gillis, Note on the transfinite diameter, *J. London Math. Soc.*, **12** (1937) 185–192.
21. M. Essén, Studies on a convolution inequality, *Arkiv f. Mat.*, **5**:9 (1963) 113–152.
22. M. Essén, Note on "A theorem on the minimum modulus of entire functions" by Kjellberg, *Math. Scand.*, **12** (1963) 12–14.
23. G. C. Evans, Potentials and positively infinite singularities of harmonic functions, *Monatsh. Math.*, **43** (1936) 419–424.
24. G. Faber, Über nach Polynomen fortschreitende Reihen, *S.B. Math.-Phys. Kl., Bayer Akad. Wiss.*, **1922**, 157–178.
25. L. Fejér, Sur les fonctions bornées et intégrables, *C.R. Acad. Sci. Paris*, **131** (1900) 984–987.
26. M. Fekete, Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten, *Math. Z.*, **17** (1932) 238–249.
27. M. Fekete, Über den transfiniten Durchmesser ebener Punktmengen, I, II, III, *Math. Z.*, **32** (1930) 108–114, 215–221; **37** (1933) 635–646.
28. W. Feller, A simple proof of renewal theorems, *Comm. Pure Appl. Math.*, **14** (1961) 285–293.
29. W. Feller and S. Orey, A renewal theorem, *J. Math. Mech.*, **10** (1961) 619–624.
30. B. de Finetti, Sul concetto di media, *Giorn. Ist. Italiano Attuari*, **2** (1931) 369–396.
31. G. Frobenius, Über die Leibnitzsche Reihe, *J. Reine Angew. Math.*, **89** (1880) 262–264.
32. O. Frostman, *Potentiel d'équilibre et capacité des ensembles avec quelques applications à la théorie des fonctions.* Thèse, Lund (1935) 118 pp.
33. G. Hamel, Eine Basis aller Zahlen und die unstetigen Lösungen der Funktionalgleichung $f(x + y) = f(x) + f(y)$, *Math. Ann.*, **60** (1905) 459–462.
34. G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge, England (1934) Cambridge Univ. Press.
35. M. Heins, On the Phragmén-Lindelöf principle, *Trans. Amer. Math. Soc.*, **60** (1946) 238–244.
36. M. Heins, Entire functions with bounded minimum modulus; subharmonic function analogues. *Ann. Math.*, (2) **49** (1948) 200–213.

37. E. Hille, *Analytic Function Theory*, I, II, Boston (1959, 1962) Ginn.
38. E. Hille, Remarks on transfinite diameters, *Gen. Topology Rel. Mod. Analysis and Algebra, Proc. Symp. Prague* (1961) pp. 211–220; Czechoslovak Acad. Sci., Prague (1962).
39. E. Hille, Green's transforms and singular boundary value problems, *J. Math. Pures Appl.*, (9) **42** (1963) 332–349.
40. E. Hille, A note on transfinite diameters, *J. Analyse Math.*, **13** (1965).
41. E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, Providence (1957) Amer. Math. Soc. Publ. 31, rev. ed.
42. W. Hoeffding, On sequences of sums of random variables. *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, **2** (1960) 213–226.
43. O. Hölder, Grenzwerte von Reihen an der Convergenzgrenze, *Math. Ann.*, **20** (1882) 535–549.
44. O. Hölder, Über einen Mittelwertsatz, *Göttinger Nachrichten*, **1889**, 38–47.
45. C. T. Ionescu Tulcea, Suboperative functions and semi-groups of operators, *Arkiv f. Mat.*, **4**:7 (1960) 55–61.
46. J. L. W. V. Jensen, Sur les fonctions convexes et les inégalités entre les valeurs moyennes, *Acta Math.*, **30** (1906) 175–193.
47. L. V. Kantorovič, Funkcionalnyi analyz i prikladnaya matematika, *Uspehi Mat. Nauk*, **3** (1945) 89–185.
48. S. Karlin, On the renewal equation, *Pac. J. Math.*, **5** (1955) 229–257.
49. B. Kjellberg, A theorem on the minimum modulus of entire functions, *Math. Scand.*, **12** (1963) 5–11.
50. K. Knopp, Zur Theorie der Limitierungsverfahren, I, II, *Math. Z.*, **31** (1929) 97–127, 276–305.
51. K. Knopp, Über die maximalen Abstände und Verhältnisse verschiedener Mittelwerte, *Math. Z.*, **39** (1935) 768–776.
52. A. N. Kolmogorov, Sur la notion de la moyenne, *Rend. Accad. Lincei*, (6) **12** (1930) 388–391.
53. K. Kunugui, Étude sur la théorie du potentiel généralisé, *Osaka Math. J.*, **2** (1950), 63–103.
54. F. Leja, Sur la définition du diamètre et de l'écart transfini d'un ensemble, *Ann. Soc. Polon. Math.*, **12** (1934) 29–34.
55. F. Leja, Sur une suite de polynomes et la represéntation conforme d'un domain plan quelconque sur le cercle, *Ann. Soc. Polon. Math.*, **14** (1936) 116–134.
56. F. Leja, Généralisation de certaines fonctions d'ensemble. *Ann. Soc. Polon. Math.*, **16** (1937) 41–52.
57. F. Leja, *Teoria Funkcji Analitycznych*, Warszawa (1957) Bibl. Mat. 14.
58. H. Minkowski, *Geometrie der Zahlen*, Leipzig, 1910 (1896) Teubner.
59. P. Montel, Sur les fonctions d'une variable réelle qui admettent un théorème d'addition algébrique, *Ann. Sci. École Norm. Sup.*, **48** (1931) 65–94.
60. P. J. Myrberg, Über stetige Funktionen einer reellen Variablen, welche ein algebraisches Additionstheorem besitzen, *J. Reine Angew. Math.*, **170** (1933) 107–122.

61. M. Nagumo, Eine hinreichende Bedingung für die Unität der Lösung der Differentialgleichung erster Ordnung, *Jap. J. Math.*, **3** (1926) 107–112.
62. M. Nagumo, Über eine Klasse der Mittelwerte, *Jap. J. Math.*, **7** (1930) 71–79.
63. F. and R. Nevanlinna, Über die Eigenschaften analytischer Funktionen in der Umgebung einer singulären Stelle oder Linie, *Acta Soc. Sci. Fennicae*, **50**:5 (1922) 46 pp.
64. W. F. Osgood, Beweis der Existenz einer Lösung der Differentialgleichung $dy/dx = f(x, y)$ ohne Hinzunahme der Cauchy-Lipschitzschen Bedingung, *Monatsh. Math.*, **9** (1898) 331–345.
65. A. Ostrowski, Sur les critères de convergence et divergence dus à V. Ermakof, *Enseignement Math.*, (2) **1** (1956) 224–257.
66. H. W. Pexider, Notiz über Funktionaltheoreme, *Monatsh. Math.*, **14** (1903) 293–301.
67. E. Phragmén, Sur une extension d'un théorème classique de la théorie des fonctions, *Acta Math.*, **28** (1904) 351–368.
68. E. Phragmén and E. Lindelöf, Sur une extension d'un principe classique de l'analyse et sur quelques propriétés des fonctions monogènes dans le voisinage d'un point singulier. *Acta Math.*, **31** (1908) 381–406.
69. G. Pólya and G. Szegö, Über den transfiniten Durchmesser (Kapazitätskonstante) von ebenen und räumlichen Punktmengen, *J. Reine Angew. Math.*, **165** (1931) 4–49.
70. M. Riesz, Intégrales de Riemann-Liouville et potentiels, *Acta Litt. Sci. Szeged*, **9** (1938) 1–42.
71. J. F. Ritt, Real functions with algebraic addition theorems, *Trans. Amer. Math. Soc.*, **29** (1927) 361–368.
72. R. A. Rosenbaum, Sub-additive functions, *Duke Math. J.*, **17** (1950) 227–247.
73. W. Schöbe, Gebrochen-iterierte Exponentialfunktion im reellen, *Z. Angew. Math. Mech.*, **38** (1958) 190–194.
74. F. Spitzer, Hitting probabilities, *J. Math. Mech.*, **11** (1962) 593–614.
75. M. H. Stone, On a theorem of Pólya, *J. Indian Math. Soc.*, **12** (1948) 1–7.
76. G. Szegö, Bemerkung zu einer Arbeit von Herrn M. Fekete: Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten, *Math. Z.*, **21** (1924) 206–208.
77. H. Terasaka, Solution of a problem of M. F. Leja, *Ann. Soc. Polon. Math.*, **23** (1950) 201–204.
78. M. Tsuji, *Potential theory in modern function theory*, Tokyo (1959) Maruzen.
79. T. Ugaheri, On the general potential and capacity, *Jap. J. Math.*, **20** (1950) 37–43.
80. E. Vincze, Eine allgemeine Methode in der Theorie der Funktionalgleichungen, 1, 2, *Publ. Math. Debrecen*, **9** (1962) 149–163, 314–323.
81. H. Wallin, On convergent and divergent sequences of equilibrium distributions, *Arkiv f. Mat.*, **4**:22 (1962) 527–549.
82. H. Wallin, Continuous functions and potential theory, *Arkiv f. Mat.*, **5**:5 (1963) 55–84.
83. N. Wiener, Tauberian theorems, *Ann. Math.*, **33** (1932) 1–100.

Addenda

84. J. B. Diaz and F. T. Metcalf, Stronger forms of a class of inequalities of G. Pólya-G. Szegö and L. V. Kantorovic, *Bull. Amer. Math. Soc.*, **69** (1963) 415–418.
85. E. V. Huntington, Sets of independent postulates for the arithmetic mean, the geometric mean, the harmonic mean, and the root mean-square, *Trans. Amer. Math. Soc.*, **29** (1927) 1–22.
86. N. D. Kazarinoff, *Analytic inequalities*, New York (1961) Holt, Rinehart and Winston.
87. R. Schimmack, Der Satz von arithmetischen Mittel in axiomatischer Begründung, *Math. Ann.*, **68** (1909) 125–132, 304.
88. P. Schweitzer, Egy egyenlötlenség az aritmetikai középértekröl [An inequality concerning the arithmetic mean], *Math. Phys. Lapok*, **23** (1914) 257–261.
89. F. F. Bonsall, *Lectures on some fixed point theorems of functional analysis*, Bombay (1962) Tata Institute of Fundamental Research, 176 pp.
90. W. Eichhorn, Lösung einer Klasse von Funktionalgleichungssysteme, *Archiv. d. Math.*, **14** (1963) 266–270.

2

Geometry

H. S. M. Coxeter

INTRODUCTION

The subject of geometry is so vast that no adequate account of its recent developments could be given in two lectures. Felix Klein, David Hilbert, or Wilhelm Blaschke might have been able to give such an account in a whole course of lectures; but that would be far beyond my powers, because there are so many branches of the subject in which I am almost as ignorant as the proverbial man in the street. I must ask you to forgive me if I concentrate on my own favorite branches, and I must take the risk of offending various geometers who will ask why I have not dealt with algebraic geometry, differential geometry, symplectic geometry, continuous geometry, metric spaces [1], Banach spaces, linear programming, and so on.

EUCLIDEAN GEOMETRY

For most of us, our first contact with geometry was in the Euclidean plane with little problems about triangles and circles. Many such problems seem prohibitively difficult to solve until we happen

to think of the right approach. A good example is the Steiner-Lehmus theorem, which says that, if two internal angle-bisectors of a triangle are equal in length, the triangle must be isosceles. This was proposed in 1840 by C. L. Lehmus and proved laboriously by Jacob Steiner. Since then about sixty proofs have been published, including one by Gilbert and MacDonnell that is surprisingly easy [34]. A purist might argue that this is an *absolute* theorem (not involving Euclid's Fifth Postulate) and should have an absolute proof. Such a proof has been given by L. M. Kelly [20, p. 16, Ex. 4].

Kakeya's problem [5, pp. 99–101; 6] asks for the plane set of least area having the "Kakeya property," which means that a unit segment can be turned continuously in the set through 180° so as to return to its original position (reversed). Besicovitch proved that there is no least area: such a set with sufficiently many holes (like a lace tablecloth) can have as small an area as we please! In view of this complete solution, it is natural to modify the problem by asking for the smallest *convex* set, or *simply connected* set, that has the Kakeya property. It was proved by Julius Pál that the smallest convex set is the equilateral triangle of unit altitude (area $1/\sqrt{3}$). It is tempting to imagine that the smallest simply connected set is bounded by the deltoid or three-cusped hypocycloid (area $\pi/8$), within which the unit segment can be moved so as to remain a tangent while both its ends run along the curve. However, a curve enclosing a considerably smaller area was discovered independently by Melvin Bloom and I. J. Schoenberg in 1963. Schoenberg picturesquely describes this star-shaped curve as resembling the locus of the end of a Foucault pendulum swinging ten thousand times. The area is less than $\pi/11$. Defining the *Kakeya constant* K to be the greatest lower bound of areas of simply connected regions having the Kakeya property, he raises the problem of computing K. So far, we know only that

$$K \leqq \frac{5 - 2\sqrt{2}}{24}\pi.$$

Fejes Tóth regards K as the case $n = 2$ of the greatest lower bound of areas of simply connected regions within which a regular n-gon can be continuously turned through $360°/n$. We know only

that this lies strictly between the area of the n-gon and the area of its circumcircle.

All three versions of Kakeya's problem can be extended to three-dimensional space, where we seek a set of least volume in which the unit segment can be moved from one direction to any other. Now even the case of the *convex* set is unsolved. Analogy suggests the regular tetrahedron of unit altitude, but Eggleston remarks that a smaller convex region will suffice.

ORDERED GEOMETRY

The name *ordered geometry* has been given by Artin [2, p. 73] to the very simple geometry whose only primitive concepts are the *point* and the relation of *intermediacy* or "betweenness." If B lies between A and C, we write $[ABC]$. The axioms [20, pp. 177–178] tell us that the relation $[ABC]$ implies $[CBA]$ but not $[BCA]$, and so on. The *segment* AB is the set of points X such that $[AXB]$, the *ray* A/B is the set of points Y such that $[BAY]$, and the *line* AB is the union of two points, two rays, and a segment:

$$A/B + A + AB + B + B/A.$$

Conceived by Pasch in 1882, this geometry was developed by Peano (1889), Hilbert (1899) and Veblen [65, Chap. I]. A model for it is easily obtained by considering the interior of a convex region in real affine or projective space. Conversely, any ordered space can be extended to a real projective space (of the same number of dimensions) by the adjunction of *ideal elements* such as "points at infinity." The best treatment of this extension is by Whitehead [62]. He, following Russell [49, p. 382], called ordered geometry "descriptive geometry," thus inviting confusion with Cayley's famous dictum "Descriptive geometry is all geometry," which really meant "Projective geometry is all geometry."

Any geometry worthy of the name has to justify itself by exhibiting interesting theorems. The first such theorem in ordered geometry is the Sylvester-Gallai theorem:

If n points are not all collinear, there is at least one line containing exactly two of them.

This was conjectured by Sylvester in 1893, revived by Karamata and Erdös in 1933, and proved (in a complicated manner) by T.

Gallai. It is always rather exciting when a problem is solved after remaining a challenge for as long as forty years. My own small contribution was to adapt Gallai's Euclidean proof to ordered geometry [20, pp. 181–182]. L. M. Kelly [20, pp. 65–66] discovered an "absolute" proof, so simple that anyone who has seen it once can always remember it.

The Sylvester-Gallai theorem and its extensions play a significant role in the part of ordered geometry that deals with *convexity* [25].

Hadwiger [35a] once asked whether every convex polyhedron can be transformed, by a finite sequence of truncations (cutting off a corner by means of a plane), into a polyhedron such that each face has a number of vertices that is a multiple of 3. The answer is yes or no according as the Four-Color Conjecture (for maps on a sphere) is true or false! This conclusion follows easily from Heawood's congruences [38, pp. 277–278; 5, pp. 230–231].

In n dimensions, a *convex polytope* can be defined *either* as the convex hull of a finite set of points *or* as a bounded set which is the intersection of finitely many half-spaces. Supporting hyperplanes of an n-dimensional convex polytope Π_n contain k-dimensional *elements* Π_k for $k = 0, 1, \ldots, n-1$. These are called *vertices, edges,* . . . , and *cells* (or *facets*). It is convenient to include also a unique n-dimensional element, the Π_n itself, and a unique (-1)-dimensional element Π_{-1}, the null set, which belongs to all the other elements. If an element Π_h belongs to an element Π_k, we say that Π_h and Π_k are *incident*. In particular, every edge Π_1 is incident with two vertices Π_0, and every Π_{n-2} is incident with two cells Π_{n-1}. More generally, whenever a Π_{k-1} and a Π_{k+1} are incident, there are just two Π_k's incident with both (for $0 \leq k \leq n-1$). In 1852 Schläfli proved that, for a polytope having N_0 vertices, N_1 edges, and so on,

$$N_0 - N_1 + N_2 - \cdots + (-1)^n N_n = 1$$

or, if you prefer,

$$-N_{-1} + N_0 - N_1 + \cdots + (-1)^n N_n = 0$$

[50, p. 190; 22, Chap. IX].

A two-dimensional polytope Π_2 is simply a polygon, which has equally many vertices and sides ($N_0 - N_1 = 0$). It follows that, in any Π_n, whenever a Π_{k-2} and a Π_{k+1} are incident, there are equally many Π_{k-1}'s and Π_k's incident with both (for $1 \leq k \leq n-1$).

If this number q_k depends only on k (that is, if it is the same for every occurrence of an incident Π_{k-2} and Π_{k+1}), the polytope is said to be (combinatorially) *regular*, and is denoted by the "Schläfli symbol"

$$\{q_1, q_2, \ldots, q_{n-1}\}.$$

For instance, $\{q\}$ is a q-gon, and the ordinary cube has the Schläfli symbol $\{4, 3\}$ because each face has four vertices (and four sides) and each vertex belongs to three edges (and three faces). Schläfli's formula (connecting the numbers N_k) is easily verified in the cases of the *simplex*

$$\{3, 3, \ldots, 3, 3\}, \qquad \text{for which} \qquad N_k = \binom{n+1}{k+1}$$

and the *cross polytope* (or "n-dimensional octahedron")

$$\{3, 3, \ldots, 3, 4\}, \qquad \text{for which}$$

$$N_k = 2^{k+1}\binom{n}{k+1} \qquad (k < n).$$

This discussion is the counterpart, in ordered geometry, of the metrical definition of a regular polytope [22, p. 288; 30, p. 134] as having a center from which all the Π_k's are equidistant, for each $k < n$.

An n-dimensional convex polytope is said to be m-neighborly if every m vertices lie in a supporting hyperplane. Thus every convex polytope is 1-neighborly, and a simplex is m-neighborly for every $m \leq n$. The possibility of other neighborly polytopes (in four or more dimensions) was recognized long ago by Brückner [8a] and Carathéodory [10a]. A 2-neighborly polytope may be described simply as having no diagonals; every two vertices are joined by an edge. Carathéodory considered a *cyclic* polytope: the convex hull of N_0 points (in affine $2m$-space) on the rational normal curve which is the locus of the point with barycentric coordinates

$$(1, t, t^2, \ldots, t^{2m}).$$

David Gale [32] has given a charmingly simple proof that such a polytope is m-neighborly and has

$$N_{2m-1} = \frac{N_0}{N_0 - m}\binom{N_0 - m}{m}$$

cells (simplexes). There is an unpublished proof by Martin

Fieldhouse that every m-neighborly polytope (with N_0 vertices in $2m$ dimensions) has this number of cells. In 1957 T. S. Motzkin conjectured that every m-neighborly polytope is combinatorially equivalent to a cyclic polytope. Gale established this conjecture for $N_0 \leq 2m + 3$. Grünbaum has pointed out that this inequality is "best possible," as one of Brückner's polytopes is the dual of a noncyclic 2-neighborly polytope having eight vertices (and twenty tetrahedral cells).

When N_{n-1} is very large, the boundary of an n-dimensional polytope resembles an $(n-1)$-dimensional tessellation or *honeycomb:* an infinite collection of cells fitting together so as to fill and cover the $(n-1)$-space, each Π_{n-2} belonging to two Π_{n-1}'s. If the Π_{n-1}'s are regular and alike, the honeycomb is said to be *regular;* it evidently has a Schläfli symbol. For instance, $\{6, 3\}$ is the familiar pattern of regular hexagons filling the Euclidean plane.

SPHERE PACKING

It seems desirable to say something about the problem of packing equal spheres in Euclidean n-space, not only because of its intrinsic interest and its connection with the geometry of numbers, but also because it has been found to have a practical application in the theory of communication.

Let M_n denote *the maximum number of spheres* (balls) *that can touch another of the same size without overlapping, in Euclidean n-space.* Everyone knows that $M_2 = 6$: seven silver dollars can be placed on a table so that one is surrounded by a ring of six. Moreover, this pattern can be extended to a packing of infinitely many circles (disks), each touching six others, and the centers of all the circles are the points of a lattice. One row of the lattice yields the analogous arrangement with $n = 1$. The one-dimensional "ball" is a line segment, and in a row of line segments each touches two others: $M_1 = 2$.

When $n = 3$, we are dealing with solid spheres, ordinary balls. In 1694 Sir Isaac Newton told David Gregory that $M_3 = 12$, but Gregory asserted that $M_3 = 13$. R. Hoppe, 180 years later, proved that Newton's value is correct. (For a fuller account of this story, see [21].) Although there is only one way of surrounding a circle with six equal circles, there are many ways of surrounding a sphere with twelve equal spheres. (This may even be done in such a way that no two of the twelve touch each other.) Two of the possible

arrangements can be continued systematically so that *each* sphere touches twelve others [41; 13]. In one of these (described by Kepler in 1611), the centers form a lattice, namely the face-centered cubic lattice [20, pp. 225, 333, 407]. We naturally call this arrangement a *lattice packing* [48a].

Let L_n denote *the number of spheres that touch each one in the densest n-dimensional lattice packing.* We know that $L_1 = M_1 = 2$, $L_2 = M_2 = 6$, $L_3 = M_3 = 12$, and $L_n \leq M_n$ always. Following Gauss [33, Vol. 1, p. 307; Vol. 2, p. 192] we can associate with each lattice a class of equivalent positive definite quadratic forms. In fact, if the lattice is generated by n vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$, its general point has the position vector

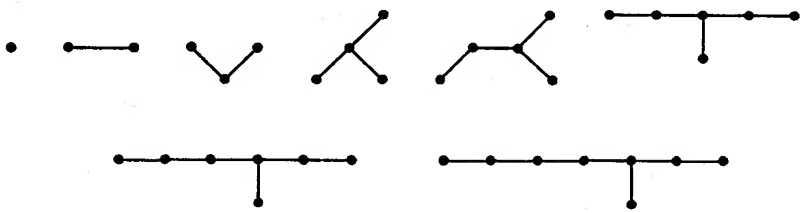
$$\Sigma x^j \mathbf{e}_j,$$

where $x^1, \ldots, x^n$ are integers. The square of the length of this vector is

$$\Sigma x^j \mathbf{e}_j \cdot \Sigma x^k \mathbf{e}_k = \Sigma\Sigma a_{jk} x^j x^k,$$

where $a_{jk} = \mathbf{e}_j \cdot \mathbf{e}_k$. We are thus led to a positive definite quadratic form; let us denote it by ϕ. The smallest positive value attained by ϕ for integers x^j is clearly equal to the square of the diameter of the largest equal spheres that can be centered at the lattice points so as to form a packing (that is, to avoid overlapping). Let us assume, for simplicity, that our spheres are of diameter 1, so that 1 is the minimal value of ϕ. Since $\det(a_{jk})$ is equal to the square of the content of the elementary cell of the lattice [20, p. 331], the densest packing occurs when this determinant is as small as possible. It follows that the search for densest packings is equivalent to the search for absolutely extreme forms [15, p. 394].

Absolutely extreme forms for $n \leq 8$ have been known since 1935. They were established for $n \leq 5$ by Minkowski [42, p. 247] and for $n = 6, 7, 8$ by Blichfeldt [7]. Forms equivalent to those of Minkowski and Blichfeldt may be represented very simply by the graphs

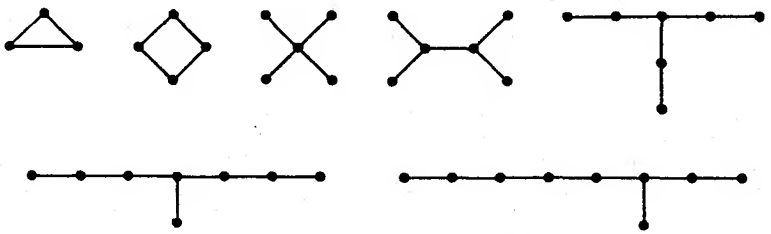


Each graph is a tree having n nodes, one for each "square" term $(x^j)^2$, and $n - 1$ branches, one for each "product" term $-x^jx^k$; thus the first three trees represent the forms

$$(x^1)^2,\ (x^1)^2 - x^1x^2 + (x^2)^2,\ (x^1)^2 - x^1x^2 + (x^2)^2 - x^2x^3 + (x^3)^2$$

[20, p. 337, Exercise, where $u^1 = x^1$, $u^2 = x^3$, $u^3 = -x^2$].

These trees will be recognized by anyone who followed the lectures of Kaplansky [40, pp. 125–126]; he used them as symbols for the Lie algebras A_1, A_2, A_3, D_4, D_5, E_6, E_7, E_8. This connection arises from the fact that every densest packing with $n \leq 8$ is symmetrical by reflection in the common tangent hyperplane at the point of contact of any two of the spheres that touch each other. Such reflections generate an infinite discrete group whose fundamental region is a simplex. When $n = 1$ the simplex is merely a line segment. The simplexes for $n = 2, \ldots, 8$ are conveniently represented by graphs



having $n + 1$ nodes, one for each bounding hyperplane (or mirror). Each branch joins two nodes that represent mirrors forming a dihedral angle of 60°. The remaining pairs of mirrors are orthogonal.

For instance, when $n = 2$ the packing consists of the incircles of the hexagons of the regular tessellation $\{6, 3\}$ (Figure 1), and the group generated by reflections in all the sides of these hexagons is adequately generated by reflections in the three sides of the shaded equilateral triangle. Two of these three sides are orthogonal to

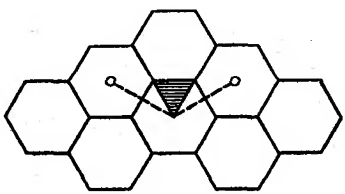


Figure 1

generating vectors of the lattice of centers. (These two vectors appear in the figure as broken lines.) Each graph in our second list can be derived from the corresponding tree in the first list by adding one extra node [15, pp. 405–412].

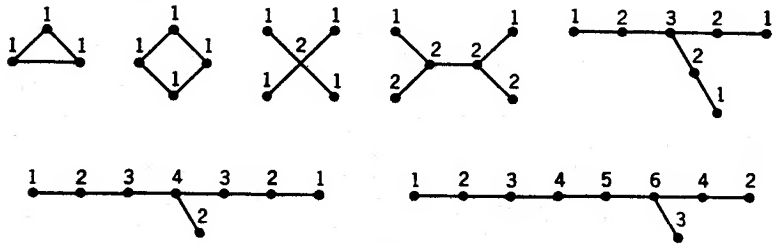


Figure 2

There is an amusingly simple way to derive the values of L_n from these augmented graphs. We place the number 1 over the extra node; then we number the remaining nodes as in Figure 2, so that the number on each is half the sum of the numbers on the adjacent nodes. (In particular, the numbers along each "leg" are in arithmetic progression.) Steinberg [54; see also 22, p. 212] has proved that L_n is equal to n times the sum of these numbers:

n	1	2	3	4	5	6	7	8
L_n	2	6	12	24	40	72	126	240

C. Muses (in a letter of December 1963) epitomized these values in the simple formula

$$L_n = n\left(\left[\frac{2^{n-2}}{3}\right] + n + 1\right) \qquad (1 \leq n \leq 8).$$

Although L_n is unknown for $n > 8$, we do know that the value given by this simple formula (for example, 468 for L_9) is too large, and that the densest lattice packing is no longer symmetrical by reflection in all the common tangent hyperplanes. It has been conjectured that $L_9 = 272$. In 1946 T. W. Chaundy believed he had a proof, but he withdrew his paper when he realized he had made the unjustifiable assumption that the densest $(n-1)$-dimensional packing must occur as a section of the densest n-dimensional packing.

Our knowledge of the numbers M_n is even less satisfactory.

We know that $M_n = L_n$ for $1 \leq n \leq 3$, and we believe that $M_4 = 24$; but it is conceivable that M_4 is 25 or 26. Obviously L_n provides a lower bound for M_n. An upper bound can be derived from Schläfli's work on the content of a simplex in spherical space (the analogue of the area of a spherical triangle) [50, p. 234; 21]. The conclusion is that

$$M_n \leq \left[\frac{2f_{n-1}(n)}{f_n(n)}\right]$$

where $f_n(x)$ is defined recursively by $f_0(x) = f_1(x) = 1$,

$$f_n(x) = \frac{1}{\pi}\int_{n-1}^{x} \frac{f_{n-2}(x-2)}{x\sqrt{x^2-1}}\,dx.$$

With the aid of an electronic computer, John Leech obtained the following values:

n	4	5	6	7	8	9	10	11	12
Bound for M_n	26	48	85	146	244	401	648	1035	1637

The most remarkable of these is the case when $n = 8$, the range of possibilities being narrowed to

$$240 \leq M_8 \leq 244.$$

The next case explains why we can be sure that $L_9 \neq 468$. For large n, the upper bound is asymptotically

$$\frac{\sqrt{\pi}}{e}\, n^{3/2} 2^{(n-1)/2}.$$

INTEGRAL QUATERNIONS AND INTEGRAL OCTAVES

The regularity of the arrangement when $n = 1$ or 2, and the near-regularity when $n = 4$ or 8, are connected with the rings of rational integers, Eisenstein integers ($a + b\rho$ where $\rho = e^{2\pi i/3}$), integral quaternions, and integral octaves. This application of geometry to the theory of numbers is, perhaps, worthy of a little more attention. Consider, for instance, the representation of the *quaternion*

$$x = x_0 + x_1 i + x_2 j + x_3 k$$

(where $i^2 = j^2 = k^2 = ijk = -1$) by the point (x_0, x_1, x_2, x_3)

in Euclidean four-space, or by the vector $\mathbf{x}$ that takes the origin to this point. In Hamilton's notation, the *conjugate* of x is

$$\bar{x} = x_0 - x_1 i - x_2 j - x_3 k$$

and the *scalar part* of x is $\mathrm{S}x = x_0 = \frac{1}{2}(x + \bar{x})$. The *norm*

$$\mathrm{N}x = x\bar{x} = \bar{x}x = {x_0}^2 + {x_1}^2 + {x_2}^2 + {x_3}^2$$

appears as the square of the length of the vector $\mathbf{x}$. Multiplication is not always commutative. The conjugate of the product xy is easily seen to be $\bar{y}\bar{x}$; therefore the conjugate of $x\bar{y}$ is $y\bar{x}$. The inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is equal to the scalar part of the quaternion product $x\bar{y}$:

$$\begin{aligned}\mathbf{x} \cdot \mathbf{y} &= x_0 y_0 + x_1 y_1 + x_2 y_2 + x_3 y_3 = \mathrm{S}(x\bar{y}) \\ &= \tfrac{1}{2}(x\bar{y} + y\bar{x}).\end{aligned}$$

If $\mathbf{a}$ and $\mathbf{b}$ are two *unit* vectors, so that $\mathrm{N}a = \mathrm{N}b = 1$, we have

$$\begin{aligned}\mathrm{N}(a + b) &= (a + b)(\bar{a} + \bar{b}) = a\bar{a} + b\bar{b} + a\bar{b} + b\bar{a} \\ &= \mathrm{N}a + \mathrm{N}b + 2\mathbf{a} \cdot \mathbf{b} = 2 + 2\cos(\mathbf{ab}),\end{aligned}$$

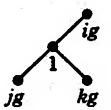
where $(\mathbf{ab})$ is the angle between the two vectors. $\mathrm{N}(a + b)$ takes the value 1 when $(\mathbf{ab}) = 120°$, and 2 when $(\mathbf{ab}) = 90°$.

Consider the reflection in the hyperplane through the origin orthogonal to the unit vector $\mathbf{a}$. Since this reflection transforms any vector $\mathbf{x}$ into $\mathbf{x} - 2(\mathbf{x} \cdot \mathbf{a})\mathbf{a}$, it transforms the corresponding quaternion x into

$$x - (x\bar{a} + a\bar{x})a = -a\bar{x}a$$

[64, p. 308].

Defining $g = \frac{1}{2}(1 + i + j + k)$, we can associate the quaternions $1, ig, jg, kg$ with the nodes of the graph



in such a way that two of them, say a and b, satisfy

$$N(a + b) = 1 \text{ or } 2$$

according as the nodes are adjacent or nonadjacent. These four quaternions generate, by addition and subtraction, the ring of

integral quaternions x whose constituents x_0, x_1, x_2, x_3 are either integers or (all of them) halves of odd integers. Also ig and jg, satisfying

$$(ig)^3 = 1, \qquad (ig)(jg)(ig) = (jg)(ig)(jg),$$

generate by multiplication the 24 unities

$$\pm 1, \ \pm i, \ \pm j, \ \pm k, \qquad \tfrac{1}{2}(\pm 1 \pm i \pm j \pm k),$$

which constitute the binary tetrahedral group of order 24 [24 (6.69)]. Finally, the four reflections

$$x \rightarrow -a\bar{x}a,$$

where $a = 1$ or ig or jg or kg, generate the group $[3^{1,1,1}]$ of order 192 [22, p. 200], which shows that the points representing the 24 unities are the vertices of the regular 24-cell 0_{111} or $\{3, 4, 3\}$, and that the lattice of points representing all the integral quaternions are the vertices of the regular four-dimensional honeycomb 1_{111} or $\{3, 3, 4, 3\}$, whose cells are 16-cells (or cross polytopes) 1_{11} or $\{3, 3, 4\}$ [27a, p. 81].

This four-dimensional representation of integral quaternions has analogues both below (in two dimensions) and above (in eight dimensions). In the subgraph

we can replace the quaternion ig (of period 3) by the complex number $\rho = e^{2\pi i/3}$. The two complex numbers 1 and ρ are a basis for the ring of Eisenstein integers, which is represented in the obvious manner by the lattice of vertices of the regular tessellation $\{3, 6\}$ of equilateral triangles. Also $-\rho$ generates the cyclic group of six unities

$$\pm 1, \ \pm \rho, \ \pm \rho^2$$

represented by a regular hexagon $\{6\}$. The hexagon is symmetrical by the reflections

$$x \rightarrow -a\bar{x}a = -a^2\bar{x},$$

where $a = 1$ or ρ. These two reflections generate the dihedral group [3], of order 6.

Octaves (often called Cayley numbers, or Cayley-Graves numbers, or the Cayley-Dickson algebra) were discovered by Graves and

Cayley, and named by Hamilton. In Dickson's notation, the octave $q + Qe$ is derived from quaternions q and Q by adjoining a new unit e which enters into (nonassociative) multiplication according to the rule

$$(q + Qe)(r + Re) = qr - \bar{R}Q + (Rq + Q\bar{r})e,$$

where the bars indicate quaternion conjugates.

The conjugate and norm of the octave

$$x = x_0 + x_1 i + x_2 j + x_3 k + x_4 e + x_5 ie + x_6 je + x_7 ke$$

are defined by

$$\bar{x} = x_0 - x_1 i - x_2 j - x_3 k - x_4 e - x_5 ie - x_6 je - x_7 ke$$

and

$$\mathrm{N}x = x\bar{x} = x_0^2 + x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 + x_6^2 + x_7^2.$$

Since

$$\overline{q + Qe} = \bar{q} - Qe = \bar{q} - e\bar{Q} = \bar{q} + \bar{e}\bar{Q}$$

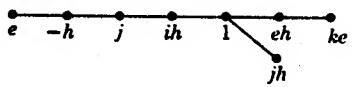
and $\mathrm{N}(q + Qe) = (q + Qe)(\bar{q} - Qe) = q\bar{q} + \bar{Q}Q = \mathrm{N}q + \mathrm{N}Q$, there is no conflict between our two uses of the bar and the N. Since $x_0 = \frac{1}{2}(x + \bar{x})$, the octave x (like the quaternion x) satisfies the quadratic equation

$$x^2 - 2x_0 x + \mathrm{N}x = 0.$$

A ring of *integral* octaves can be selected in seven ways. One way uses the basis

$$1, j, e, ke, h, ih, jh, eh,$$

where $h = \frac{1}{2}(i + j + k + e)$ [12, p. 569]. Our rule about $\mathrm{N}(a + b)$ applies again to the graph



We represent the octave

$$x_0 + x_1 i + x_2 j + x_3 k + x_4 e + x_5 ie + x_6 je + x_7 ke$$

by the vector or point $(x_0, x_1, \ldots, x_7)$ in Euclidean eight-space. The reflection that reverses a unit vector a is still given by

$$x \to -a\bar{x}a.$$

(Although multiplication is not universally associative, we do have $ab \cdot a = a \cdot ba$, so that such an expression as aba is unambiguous.)

The four octaves i, j, e, h generate by multiplication the loop of 240 unities

$$\pm 1,\ \pm i,\ \pm j,\ \pm k,\ \pm e,\ \pm ie,\ \pm je,\ \pm ke,$$

$$\tfrac{1}{2}(\pm 1 \pm ie \pm je \pm ke), \qquad \tfrac{1}{2}(\pm i \pm j \pm k \pm e),$$

$$\tfrac{1}{2}(\pm 1 \pm j \pm k \pm ie), \qquad \tfrac{1}{2}(\pm i \pm e \pm je \pm ke), \text{ etc.}$$

(where "etc." means that we may cyclically permute i, j, k). Finally, the eight reflections

$$x \rightarrow -a\bar{x}a,$$

where a takes the values $1, j, e, ke, h, ih, jh, eh$, generate the group $[3^{4,2,1}]$ of order $192 \cdot 10!$ [22, p. 204], which shows that the points representing the 240 unities are the vertices of the semiregular polytope 4_{21}. (These numbers 4, 2, 1 indicate the lengths of the three legs of the graph.) Also the lattice of points representing all the integral octaves are the vertices of the eight-dimensional honeycomb 5_{21}, whose cells are simplexes 5_{20} and cross polytopes 5_{11}.

Since a cross polytope of edge 1 has circumradius $\sqrt{\frac{1}{2}}$ while a simplex has a smaller circumradius, every point in the space is within a distance $\sqrt{\frac{1}{2}}$ of some vertex of the honeycomb. It follows that, *for every octave* ξ, *there is an integral octave* x *such that*

$$\mathrm{N}(\xi - x) \leq \tfrac{1}{2}$$

[12, p. 577]. Similar considerations yield the same result for a quaternion ξ and an integral quaternion x, and for a complex number ξ and a Gaussian integer x. However, since the circumradius of an equilateral triangle is only $\sqrt{\frac{1}{3}}$, for every complex number ξ there is an Eisenstein integer x such that $\mathrm{N}(\xi - x) \leq \frac{1}{3}$. (Here we are using the word "norm" in Hamilton's sense. Workers in Hilbert space use the same word for $|\xi - x|$, the square root of Hamilton's norm.) It is possible that this geometric approach to algebraic numbers could be further exploited.

PROJECTIVE GEOMETRY

In projective geometry the primitive concepts are the *point*, the *line*, and the relation of *incidence*. The axioms [23, p. 15] tell

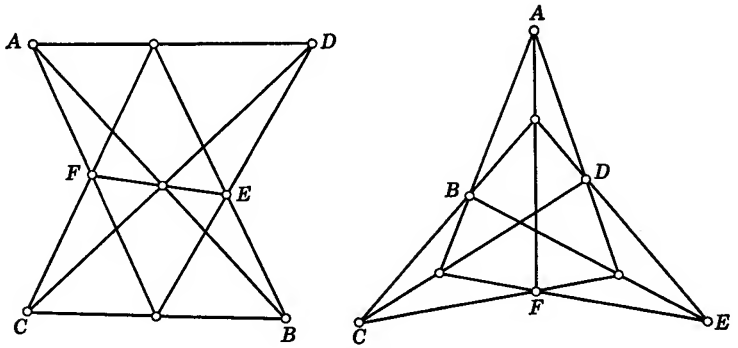


Figure 3

us that if AB meets CD (and the four points are all distinct) then AC meets BD, and so on. If C is not on AB, we can define the *plane* ABC as consisting of all the points on all the lines through C that meet AB, and all the lines that join pairs of these points. If we include an axiom saying that, for any plane ABC, there is a point D not in this plane (so that the number of dimensions is greater than 2), we can prove Desargues's theorem about perspective triangles and deduce the uniqueness of the fourth point of a harmonic set. If, on the other hand, we assume that all points lie in one place (so that the number of dimensions is 2), we are free to assert or deny Desargues's theorem, that is, to work in a *Desarguesian* plane or a *non-Desarguesian* plane, respectively. A plane may be non-Desarguesian and still admit a unique fourth harmonic point for any three given collinear points; it is then called a Moufang plane [36, p. 370; 46, p. 137].

Another famous theorem that may be either taken as an axiom or denied is Pappus's theorem (Figure 3) which says, of the lines joining six coplanar points, that if AB, CD, EF are concurrent and DE, FA, BC are concurrent, then also AD, BE, CF are concurrent [23, p. 90]. (G. Hessenberg's proof that Pappus implies Desargues has been published in many versions [e.g., 16, pp. 1, 46], but almost every author disregards at least one exceptional situation. Apparently the first complete proof is that of Pickert [47, pp. 144–148].) Since it is possible to deny Pappus without denying Desargues, geometries without Pappus can occur in any number of dimensions.

Pappus's theorem is equivalent to the Fundamental Theorem,

which says that a projectivity is determined by its effect on three collinear points. This plays an essential role in the classical treatment of polarities and conics [23, pp. 41–89]. In a projective plane satisfying Pappus's theorem, a polarity is an involutory correspondence between points and lines, such that each range corresponds to a projectively related pencil. A point A (or B) and the corresponding line a (or b) are called *pole* and *polar*. If B lies on a, then A lies on b, A and B are called *conjugate points*, a and b are called *conjugate lines*. If a self-conjugate point exists, it lies on a *conic* whose points and tangents are precisely all the self-conjugate points and self-conjugate lines. Other points are said to be *exterior* or *interior* according as they lie on two tangents or on no tangent. The polars of such points are *secants* or *nonsecants*, respectively. On a secant AB, any pair of conjugate points is a pair of harmonic conjugates with respect to A and B. (These definitions and deductions, which are commonplace for the real projective plane, remain valid for the general Desarguesian plane, and can be adapted to non-Desarguesian planes [4].)

In the presence of a given polarity, a point Q is said to be *accessible* from a point P if Q is the harmonic conjugate of P with respect to some pair of distinct points which are conjugate for the polarity. The relation of accessibility is evidently reflexive and symmetric. By an application of Hesse's theorem, it is also transitive [23, pp. 101, 148], and the points of the plane can be distributed into one or more classes of mutually accessible points.

If the polarity admits self-conjugate points, their locus (the conic) constitutes a single class, and the exterior points constitute another single class. Thus the exterior points on a secant form a segment according to the ingenious definition of Pieri [16, p. 38]. Over some fields (such as real, rational, or finite) there are also interior points (lying on no tangent). We shall see later on that, in any real or finite plane, the interior points constitute a single class; but in the rational plane the interior points are distributed among infinitely many classes.

One of the most elegant developments in projective geometry has been the synthetic coordinatization of the general projective plane. This began in 1847 with von Staudt's calculus of "throws," and continued with Hilbert's remark that, in any Desarguesian plane, suitable definitions of addition and multiplication exhibit the points of a range as a division ring or "skew field" or *corpus*,

which becomes a (commutative) *field* when Pappus's theorem holds. Further investigation by Oswald Veblen and J. H. M. Wedderburn inspired Marshall Hall [36, p. 123] to invent the *ternary ring* which enabled him to coordinatize a non-Desarguesian plane. This exciting story has been told many times, for example, by Pedoe [46, Chap. VI].

The most interesting case is that of the Moufang plane, in which the ternary ring reduces to an *alternative division ring:* addition is an Abelian group, both distributive laws hold, and each $a \neq 0$ has an inverse a^{-1} satisfying

$$a^{-1}a = aa^{-1} = 1, \qquad a^{-1}(ab) = b = (ba)a^{-1},$$

whence $a(ab) = (aa)b$, $(ba)a = b(aa)$. It has been proved by Bruck and Kleinfeld [9] that every nonassociative ring of this kind is simply the ring of octaves over a field. (Earlier, we tacitly assumed this field to consist of the real numbers.) This aspect of the Moufang plane has been developed intensively by Freudenthal [31] and van der Blij [8].

In the algebraic approach to n-dimensional projective geometry over an arbitrary corpus [53, Part II], a point (x) and a hyperplane $[X]$ are defined to be sets of "equivalent" symbols

$$(x_1, \ldots, x_{n+1}) \qquad \text{and} \qquad [X_1, \ldots, X_{n+1}],$$

where the x's and X's are not all zero, the rules for equivalence are that $(x_1, \ldots, x_{n+1})$ is equivalent to $(x_1\lambda, \ldots, x_{n+1}\lambda)$ $(\lambda \neq 0)$ and $[X_1, \ldots, X_{n+1}]$ to $[\mu X_1, \ldots, \mu X_{n+1}]$ $(\mu \neq 0)$, and the condition for incidence is $\{Xx\} = 0$, where

$$\{Xx\} = X_1x_1 + \cdots + X_{n+1}x_{n+1}.$$

It follows that the points collinear with (x) and (y) are $(x\lambda + y\mu)$, meaning $(x_1\lambda + y_1\mu, \ldots, x_{n+1}\lambda + y_{n+1}\mu)$. In particular, any three distinct collinear points may be expressed as

$$(x), (y), (x + y).$$

If a fourth point on the same line is $(x\lambda + y)$, the number λ is called the *cross ratio* of the four collinear points [46, p. 170]. Dually, μ is the cross ratio of four "coaxial" hyperplanes

$$[X], [Y], [X + Y], [\mu X + Y].$$

If the four points are sections of the four hyperplanes, we have

$$\{Xx\} = 0, \qquad \{Yy\} = 0, \qquad \{Xy\} + \{Yx\} = 0$$

and

$$\mu\{Xy\} + \{Yx\}\lambda = 0, \qquad \text{whence} \qquad \mu\{Xy\} = \{Xy\}\lambda.$$

Thus we can conclude $\lambda = \mu$ if and only if the corpus is a field. In other words, the commutative law is logically equivalent to the invariance of cross ratio under projectivities, to the fundamental theorem of projective geometry, to Pappus's theorem [23, p. 38], to Möbius's theorem about mutually inscribed tetrahedra [20, p. 258] and to Gallucci's eight-line theorem [14, pp. 444–445]:

If three skew lines all meet three other skew lines, any transversal to the first set of three meets any transversal of the second set.

In the commutative case, let $\{AB, CD\}$ denote the cross ratio λ of four collinear points A, B, C, D, so that, if A is (x) and B is (y) and C is $(x + y)$, then D is $(x\lambda + y)$, which is now the same as $(\lambda x + y)$. It follows easily [46, p. 171] that

$$\{AB, DC\} = \lambda^{-1} \qquad \text{and} \qquad \{AC, BD\} = 1 - \lambda.$$

Also, if C and D lie in hyperplanes α and β, respectively, we write $\{AB, CD\} = \{AB, \alpha\beta\}$, $\{AC, BD\} = \{A\alpha, B\beta\}$, and so on. If α is $[X]$ and β is $[Y]$, so that $(x + y)$ lies on $[X]$, and $(x\lambda + y)$ on $[Y]$, we have $\{Xx\} + \{Xy\} = 0$ and $\{Yx\}\lambda + \{Yy\} = 0$, whence

$$\{AB, \alpha\beta\} = \frac{\{Xx\}\{Yy\}}{\{Yx\}\{Xy\}}$$

[39, pp. 120, 136].

According to a famous theorem of Wedderburn, every *finite* corpus is commutative (that is, every finite corpus is a Galois field). All known proofs, such as that of Ernst Witt [63], are deeply algebraic. Geometers are challenged by the thought that there may be a nonalgebraic proof of the equivalent statement that Pappus's theorem (or Möbius's or Gallucci's) holds in every finite Desarguesian plane. Such a thought was doubtless the motivation for the large section of Segre's book [53] that is devoted to line geometry over a general corpus.

If a, b, c are three skew lines, the *regulus abc* has *directrices* which are their transversals, and *generators* which are transversals of triads of directrices. If the corpus is not a field, Gallucci's theorem

fails, and so there are more generators than directrices. Such considerations provide a basis for the noncommutative theory of *conics*, which can be defined as sections of reguli.

One of the earliest instances of a non-Desarguesian plane was invented by Moulton [43]. However, that has an unpleasant air of artificiality. Far more satisfying is the example of Segre [51, p. 39], where the points and lines of a non-Desarguesian projective plane are represented by the points of the real projective plane and a certain two-parameter family of cubic curves, namely

$$(X_1^2 + X_2^2 + X_3^2)(X_1x_1 + X_2x_2 + X_3x_3)(x_1^2 + x_2^2 + x_3^2) + \lambda X_3^3 x_3^3 = 0.$$

where the X's are homogeneous parameters, varying from curve to curve, and λ is a sufficiently small constant.

Before 1892, coordinates were always either real or complex. Then Fano [29] described an n-dimensional geometry in which the coordinates belong to the field of residue-classes modulo a prime p, so that the number of points on a line is $p + 1$. In 1906 Veblen and Bussey [61] gave this finite Projective Geometry the name $PG(n, p)$ and extended it to $PG(n, q)$ where the coordinates belong to the Galois field $GF(q)$, q being any power of a prime. Many of the properties of this geometry depend only on the number $q + 1$ of points on a line. For instance [53, p. 27], only the most basic axioms of incidence are needed to prove that the number of k-dimensional subspaces S_k in the n-space S_n is

$$\frac{(q^{n+1} - 1)(q^n - 1) \ldots (q^{n+1-k} - 1)}{(q^{k+1} - 1)(q^k - 1) \ldots (q - 1)} = \frac{(q^{n+1} - 1)(q^n - 1) \ldots (q^{k+2} - 1)}{(q^{n-k} - 1)(q^{n-k-1} - 1) \ldots (q - 1)}.$$

The former expression is appropriate when $k \leq \frac{1}{2}(n - 1)$, the latter when $k \geq \frac{1}{2}(n - 1)$. In particular, the finite plane contains $q^2 + q + 1$ points and (of course) the same number of lines.

In 1901 Dickson [26] represented many finite groups, especially simple groups, in finite projective spaces. About 1940 Brauer mentioned the desirability of finding a geometric representation for the Mathieu groups M_{11}, M_{12}, M_{22}, M_{23}, M_{24}. This remark inspired my work [18] which was continued by Todd [60], with the conclusion that the quintuply transitive groups M_{12} and M_{24} can

be represented as collineation groups in $PG(5, 3)$ and $PG(11, 2)$, respectively.

ELLIPTIC AND HYPERBOLIC POLARITIES

In the projective plane over a (commutative) field, a polarity is given by a bilinear form

$$(xy) = \Sigma\,\Sigma\, c_{jk}x_jy_k, \qquad c_{jk} = c_{kj}, \qquad \det(c_{jk}) \neq 0.$$

Two points (x) and (y), meaning (x_1, x_2, x_3) and (y_1, y_2, y_3), are conjugate if and only if $(xy) = 0$. Since the harmonic conjugates

$$(x \pm \mu y) = (x_1 \pm \mu y_1, x_2 \pm \mu y_2, x_3 \pm \mu y_3) \qquad (\mu \neq 0)$$

are conjugate (for the polarity) if

$$(xx) = \mu^2(yy),$$

this relation (for some nonzero square μ^2) is the condition for (x) to be accessible from (y) [23, pp. 124, 153].

If the quadratic form (xx) is indefinite, self-conjugate points exist, and their locus is the conic

$$(xx) = 0.$$

Multiplying the c_{jk} (if necessary) by a suitable constant, we can ensure that, for some particular exterior point (y), (yy) is a square. Then, since the exterior points constitute a single class of mutually accessible points, (xx) is a nonzero square for every exterior point (x) and a nonsquare for every interior point (x) [23, pp. 126, 155; 45, p. 201]. For instance, if

$$(xy) = x_1y_1 + x_2y_2 - x_3y_3,$$

so that $(xx) = {x_1}^2 + {x_2}^2 - {x_3}^2$, the conic $(xx) = 0$ has tangents $x_2 \pm x_3 = 0$, and we can take (y) to be their point of intersection $(1, 0, 0)$.

Over the real field, the statement "(xx) is a nonzero square" becomes simply "$(xx) > 0$."

Over the complex field, every number is a square, so there are no interior points. Over the real field or any finite field, the product (or quotient) of two nonsquares is a square, and so *the interior points constitute a single class.* Over the rational field, we can find

points (x) for which (xx) takes in turn various prime values; such points are interior but inaccessible from one another.

The same distinction between real and rational fields occurs when the polarity has no self-conjugate points, that is, when all the points of the plane are "interior." If such an "elliptic" polarity could exist over a finite field (of odd characteristic), that is, in a finite plane $PG(2, q)$ (q odd), the number of points in each class of accessible points would be $\frac{1}{2}(q^2 + 1)$, which is not a divisor of $q^2 + q + 1$, the number of points in the whole plane [3, pp. 123–124; 23, pp. 101, 149]. Therefore *finite geometries* of two (or more) dimensions *admit no elliptic polarities* [2, p. 144; 52, p. 4].

This result is equivalent to the case $n = 3$ of the following theorem due to Chevalley [11]:

Every Galois field is quasi-algebraically closed.

This means that, over a finite field, if $F(x_1, \ldots, x_n)$ is any homogeneous polynomial of degree less than n, the equation $F = 0$ has a nontrivial solution.

CONICS IN THE REAL PLANE

The theory of accessibility shows that, in the *real* projective plane, two points on a conic decompose the secant joining them into two segments, consisting of exterior and interior points, respectively. The interior segment is naturally called a *chord*. Figure 4 shows two points A and B on a chord MN, so arranged that A and M

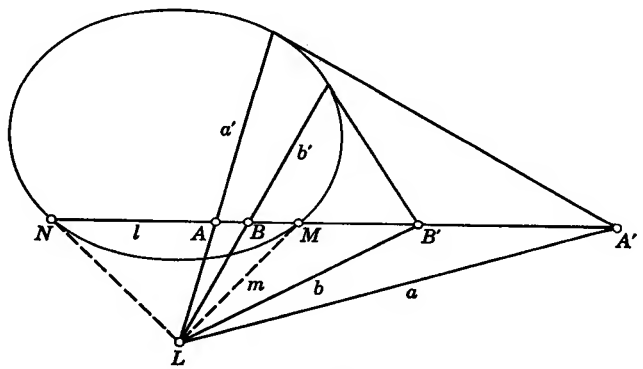


Figure 4

separate B and N. The polars of A and B, say a and b, join L, the pole of the line MN, to points A' and B' on this line. These last points are the harmonic conjugates of A and B with respect to M and N [23, p. 73]. Hence, if M, N, A, B are

$$(x),\ (y),\ (x + y),\ (\lambda x + y),$$

respectively, then A' is $(-x + y)$ and B' is $(-\lambda x + y)$. Since the cross ratio of $(\kappa x + y)$, $(\lambda x + y)$, $(\mu x + y)$, $(\nu x + y)$ is

$$\frac{(\kappa - \mu)(\lambda - \nu)}{(\kappa - \nu)(\lambda - \mu)}$$

[23, pp. 119, 152], it follows that

$$\{AB, ba\} = \{AB, B'A'\} = \frac{(1 + \lambda)(\lambda + 1)}{(1 + 1)(\lambda + \lambda)} = \frac{(\lambda + 1)^2}{4\lambda},$$

where $\lambda = \{MN, AB\} = \{AB, MN\}$. This result extends easily to real projective n-space, with a nonruled quadric instead of a conic: if α and β are the polar hyperplanes of two points A and B on a chord MN, then

$$\{AB, \beta\alpha\} = \frac{(\lambda + 1)^2}{4\lambda},$$

where $\lambda = \{AB, MN\}$.

CONICS AND k-ARCS IN A FINITE PLANE

In any *finite* projective plane, a set of k points, no three collinear, is called a *k-arc*. A k-arc is said to be *complete* if it is not part of a $(k + 1)$-arc. Segre [53, pp. 270–294] has proved that, in a Desarguesian plane $PG(2, q)$ where q is odd, every $(q + 1)$-arc is complete, the $(q + 1)$-arcs are just the conics, and there are no complete q-arcs. But complete $(q - 1)$-arcs occur when $q = 7$, 9, 11, 13; also complete 6-arcs and 7-arcs occur when $q = 9$.

In a plane $PG(2, q)$ where q is even (namely a power of 2), the $q + 1$ tangents of a $(q + 1)$-arc concur in a point called the *nucleus* (or "center") of the $(q + 1)$-arc. The $(q + 1)$-arc is incomplete, but yields a complete $(q + 2)$-arc when its nucleus is added. Each point of the $(q + 2)$-arc is the nucleus of the $(q + 1)$-arc formed by the remaining points, but these $q + 1$ points usually *do not form a*

conic! It has been proved with the aid of an electronic computer that, when $q = 16$, some $(q + 2)$-arcs contain no conic. The same situation arises when $q = 32$ or 128 or any higher power of 2, but the case when $q = 64$ remains undecided.

In virtue of Wedderburn's theorem, the only possible coordinates for a finite non-Desarguesian plane belong to a ternary ring which is not a corpus. Still denoting the number of points on a line by $q + 1$, we naturally ask what are the possible values for q. Although some progress has been made, this remains a challenging problem. The number q is called the *order* of the plane.

The painstaking work of Tarry [59] on Euler's Officers' Problem [5, p. 190] proves the impossibility of $q = 6$. It is known [37] that the only planes of other orders less than 9 are the Desarguesian planes $PG(2, q)$; but non-Desarguesian planes are known for $q = 9$ and many greater values, including p^r with p odd and $r \geq 2$. In every known case, q is a power of a prime [36, p. 394]. Bruck and Ryser [10] proved that there cannot be a plane with $q \equiv 1$ or 2 (mod 4) unless q is expressible in the form $a^2 + b^2$. Since 10 is so expressible, but 6 not, this result includes Tarry's but leaves open the possibility of $q = 10$.

In $PG(2, q)$, where q is a power of 2, every complete quadrangle has collinear diagonal points. Conversely, this property for every complete quadrangle in a finite plane is sufficient to make the plane Desarguesian [35]. In $PG(2, q)$, where q is odd, every complete quadrangle has noncollinear diagonal points. Hanna Neumann [44] has conjectured that this property for every complete quadrangle suffices to make the finite plane Desarguesian.

HYPERBOLIC GEOMETRY

When an ordered space is extended to a real projective space (that is, imbedded in a real projective space), each ordered line is extended to a projective line (on which the order of points is not serial but cyclic). On such a line there is either a single ideal point or an interval of ideal points. If every line has a single ideal point (its *point at infinity*) the ordered space is *affine* [20, Chap. 13]: through any point there is a unique line parallel to a given line, and we can compare distances along parallel lines. If every line has an interval of ideal points, the ordered space is *hyperbolic*, the ends of the ideal segments are called points at infinity (or simply *ends*),

and their locus (in the projective space) is a nonruled quadric [17, p. 195] or, in two-dimensional case, a conic. This is Klein's approach to non-Euclidean geometry: the points and lines are the interior points and chords of a nonruled quadric (or conic).

When the points and lines of the hyperbolic plane are regarded as the interior points and chords of a conic in the real projective plane, the isometries of the former appear in the latter as collineations leaving the "absolute" conic invariant [17, pp. 201–203]. In the projective plane, a collineation of period 2 is a harmonic homology [23, p. 55], and this leaves a conic invariant if the center and axis are pole and polar. When applied to the absolute conic, this collineation appears as a reflection or a half-turn according to the nature of its axis. If the axis is a secant (so that its pole is exterior) the homology is the reflection in this line. If the axis is a nonsecant, so that its pole is interior, the homology is the half-turn about this point. It follows that *perpendicular* lines are conjugate with respect to the absolute conic. The various lines perpendicular to a given line are ultraparallel [55] to one another: their common point, being the pole of the given line, is exterior, and thus does not belong to the hyperbolic plane itself.

Any two *ends*, M and N, determine a line $l = MN$ whose pole L is the intersection of the tangents at M and N (Figure 5). For any point P, not on l, the segments PM and PN are the rays from P *parallel* to l (in the sense of Gauss, Bolyai, and Lobachevsky): they do not meet l (in ordinary points), but every ray within the angle NPM does meet l. This angle is bisected by the line PL, which, being perpendicular to l, acts as a mirror reflecting PM into PN. This angle-bisecting technique, combined with continuity, soon yields a measurement of angles. In particular, $\angle LPM$ is the *angle of parallelism* for the point P and line l.

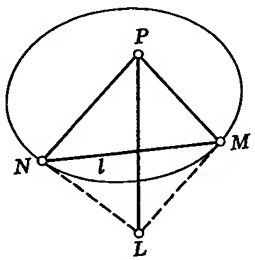


Figure 5

The measurement of *distance* can be developed from the observation that two directed segments, AB and $A'B'$, on chords MN and $M'N'$ of the absolute conic, are congruent if $ABMN \overline{\wedge} A'B'M'N'$, that is, if

$$\{AB, MN\} = \{A'B', M'N'\}.$$

The essential property of a directed distance is its additivity: if $[ABC]$, then

$$AB + BC = AC.$$

Taking into consideration the relation

$$\{AB, MN\}\{BC, MN\} = \{AC, MN\}$$

[17, p. 77], which holds for any five collinear points, we see that this requirement is neatly satisfied by defining

$$AB = \tfrac{1}{2} \log \{AB, MN\}.$$

(The $\frac{1}{2}$ is inserted for convenience [17, p. 201].) Now let a and b be the absolute polars of A and B, as in Figure 4. Setting $\lambda = e^{2AB}$ in our formula

$$\{AB, ba\} = \frac{(\lambda + 1)^2}{4\lambda}, \qquad \text{where} \qquad \lambda = \{AB, MN\},$$

we deduce

$$\sqrt{\{AB, ba\}} = \frac{e^{2AB} + 1}{2e^{AB}} = \frac{e^{AB} + e^{-AB}}{2} = \cosh AB.$$

Thus the distance AB is given by

$$\cosh AB = \sqrt{\{AB, ba\}}, \qquad \operatorname{sech} AB = \sqrt{\{AB, ab\}},$$
$$\tanh AB = \sqrt{\{Aa, Bb\}}.$$

Similarly in n dimensions, if α and β are the absolute polar hyperplanes of A and B,

$$\cosh AB = \sqrt{\{AB, \beta\alpha\}} \qquad \text{and} \qquad \tanh AB = \sqrt{\{A\alpha, B\beta\}}.$$

In terms of projective (homogeneous) coordinates, let the absolute quadric have the equation $(xx) = 0$, where

$$(xy) = \Sigma\Sigma c_{jk} x_j y_k, \qquad c_{jk} = c_{kj}, \qquad \det (c_{jk}) \neq 0.$$

Then the polar hyperplane of the point (y) is $[Y]$, where

$$Y_j = \Sigma c_{jk} y_k.$$

Let A, B, α, β be (x), (y), $[X]$, $[Y]$, so that

$$\{Yx\} = \Sigma Y_j x_j = (xy).$$

Then

$$\cosh^2 AB = \{AB, \beta\alpha\} = \frac{\{Yx\}\{Xy\}}{\{Xx\}\{Yy\}} = \frac{(xy)^2}{(xx)(yy)}.$$

Somewhat similar considerations [17, p. 225] yield the formula

$$\cos^2 (\alpha\beta) = \frac{(xy)^2}{(xx)(yy)}$$

for the angle $(\alpha\beta)$ between two intersecting hyperplanes, α and β, whose poles are exterior points (x) and (y) on a nonsecant.

EXTERIOR-HYPERBOLIC GEOMETRY

As long ago as 1907, Study [55] described the geometry of the points *exterior* to a nonruled quadric $(xx) = 0$ in real projective n-space, that is, of the poles of the hyperplanes in hyperbolic n-space. He observed that the join of two such points may be either a secant or a tangent or a nonsecant. In the case of the secant, we can simply interchange the roles of the primed and unprimed letters in Figure 4, and conclude, as before, that

$$\cosh^2 AB = \frac{(xy)^2}{(xx)(yy)}.$$

In the case of the nonsecant, we borrow a trick from spherical trigonometry and measure the distance AB by the angle $(\alpha\beta)$ between the polar hyperplanes, so that

$$\cos^2 AB = \frac{(xy)^2}{(xx)(yy)}.$$

In the intermediate case (when the line AB is a tangent), we have

$$(xy)^2 = (xx)(yy)$$

[23, pp. 126, 154, Ex. 6]; so it is natural to define the length of AB to be zero.

Dividing the coefficients c_{jk} by the constant (yy), where (y) is a particular exterior point, we can contrive (as in our section on "Elliptic and Hyperbolic Polarities") to make (xx) positive for *every* exterior point. Then, dividing the coordinates of each point (x) by $\sqrt{(xx)}$, we can make $(xx) = 1$ for every exterior point [17, p. 281]. In other words, we replace our homogeneous coordinates by redundant nonhomogeneous coordinates satisfying $(xx) = 1$, with the pleasing result that

$$\cosh AB = |(xy)|, \qquad AB = 0, \qquad \text{or} \qquad \cos AB = |(xy)|$$

according as the line AB is a secant, a tangent, or a nonsecant.

For instance, as we have already seen, $(1, 0, 0)$ is exterior to the conic

$$x_1^2 + x_2^2 - x_3^2 = 0,$$

so we can assume $x_1^2 + x_2^2 - x_3^2 = 1$ for every point (x) in the exterior-hyperbolic plane determined by this conic. Study observed [55, p. 108] that in a triangle whose sides are lines of the first kind (secants), one side is *longer* than the sum of the other two!

RELATIVITY

In its geometric aspect, the space-time of relativity theory is a diagram consisting of a four-dimensional space whose points represent events. As Synge remarks [57, pp. 5, 6]: "Anything that happens is an *event*, but we sharpen the concept to mean an occurrence that takes up no room and has no duration. . . . All possible events form a four-dimensional continuum." The locus of events in the "life" of any particular particle, being one-dimensional, is a curve in the four-space. Following Minkowski [42a, p. 55], we call this curve the *world line* of the particle. Synge's most charming and revealing description of world lines was given in a short paper in the *New Scientist* [56].

The passage of time appears in the diagram as a direction along the world line of each particle: thus the world line is a *directed* curve. Each event in the history of the particle appears as a point decomposing the curve into two parts: the past and the future. In particular, the world line of an unaccelerated particle appears as a straight line, on which the past and future are two opposite rays.

In the familiar spaces (Euclidean, affine, projective, non-Euclid-

ean) all positions are alike and all directions are alike. In the four-dimensional space-time, all positions are alike (provided we ignore the "mountain ranges" that represent matter) but directions are of five essentially different kinds. For each event O, there are the directions of possible future events of a particle at O, the directions of possible past events of such a particle, and the directions of events P which are neither before nor after O. These three sets of directions are separated by the two nappes of the *light cone* (or "null cone," or "isotropic cone") whose generators are the world lines of light signals (or photons) that could be emitted from O or received at O [57, p. 41]. This cone is, of course, a three-dimensional pseudo-manifold which joins its vertex O by straight lines to all the points on a nonruled quadric surface like the familiar sphere.

The word *cone* is appropriate not merely because it gives a rough idea of the light rays associated with any event but because the most natural geometric spaces to use for our four-dimensional diagram are *real affine four-space* and *real projective four-space*, in both of which a quadric cone can be defined and a distinction can be made between three sets of lines through the vertex: generators that lie entirely on the cone, exterior lines on pairs of tangent hyperplanes, and interior lines that lie on no tangent hyperplanes. A generator is called a *null* line (or an *isotropic* line) because the measurement of any interval along it is zero. Other lines are said to be *timelike* or *spacelike* according as they are interior or exterior. The world line of a photon is a null line. The world lines of any other unaccelerated particle is timelike. The join OP of two events that are essentially incapable of influencing each other is spacelike [57, p. 19; 58, p. 108].

Let A, B, C be three unaccelerated particles, initially together and remaining collinear but parting with relative velocities v_{BC}, $v_{CA} = -v_{AC}$, and v_{AB}. In nonrelativistic kinematics we would have

$$v_{AC} = v_{AB} + v_{BC}.$$

The special theory of relativity changes this to

$$v_{AC} = (v_{AB} + v_{BC})/(1 + v_{AB}v_{BC}/c^2),$$

where c is the velocity of light. The addition rule

$$\tanh (x + y) = (\tanh x + \tanh y)/(1 + \tanh x \tanh y)$$

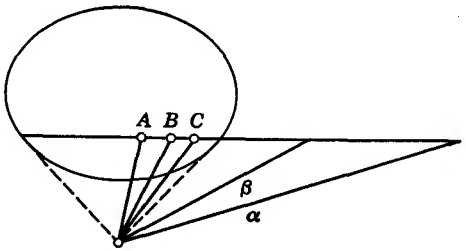


Figure 6

suggested to Robb the idea of using, instead of the velocity v, the *rapidity* r, defined by $v = c \tanh r$, so that

$$r_{AB} + r_{BC} = r_{AC}$$

[28, p. 22; 57, p. 126].

Robb showed in his Appendix [48, p. 405] that the timelike directions at any event O can be represented by the points of a hyperbolic three-space in such a way that rapidities appear as distances. Thus the directions of the world lines of our three particles are represented by three collinear points A, B, C (Figure 6) and the equation $r_{AB} + r_{BC} = r_{AC}$ becomes

$$AB + BC = AC.$$

It follows that the null vectors through O are represented by the "points at infinity" of the hyperbolic space, and the light cone at O is represented by the locus of points at infinity, which is the absolute quadric: a nonruled quadric in real projective three-space. From this it is a natural step to complete the picture by using the points outside the quadric to represent the spacelike directions from O.

Using the projective formula for a hyperbolic distance, we find that, if α and β are the polar planes of A and B, as in Figure 6, the relative rapidity $r_{AB} = AB$ of these two particles is given by

$$\tanh AB = \sqrt{\{A\alpha, B\beta\}}.$$

Therefore their relative *velocity* is simply

$$v_{AB} = c \tanh AB = c\sqrt{\{A\alpha, B\beta\}}:$$

a pretty connection between kinematics and projective geometry!

Although Robb [48, p. 22] was aware of the fact that this way of

representing the directions of world lines is valid in the general theory of relativity, he was chiefly concerned with the special theory, according to which the underlying four-space is affine and the three kinds of directions from any event are *parallel* to those from any other. He gave a sufficient set of axioms to describe a Minkowskian space-time which, from our standpoint, is simply a real affine four-space with a pseudo-Euclidean metric. In its projective aspect, this is a projective four-space in which one hyperplane has been specialized and called the three-space at infinity. Two lines are parallel if and only if their points at infinity coincide. The metric is imposed by specializing a nonruled quadric surface Ω in this projective three-space. Two lines are orthogonal if and only if their points at infinity are conjugate with respect to Ω. At any event O, the light cone joins O to Ω. The timelike and spacelike lines join O to the interior and exterior points. In fact, this quadric Ω, which is a common section of all light cones, serves as the absolute for Robb's hyperbolic space whose ordinary points represent the timelike vectors.

Using homogeneous coordinates $x_1, \ldots, x_5$ in the projective four-space, we are free to choose the simplex of reference and "unit point" so that the special hyperplane is $x_5 = 0$ and the quadric surface Ω is

$$x_1^2 + x_2^2 + x_3^2 - x_4^2 = 0, \qquad x_5 = 0.$$

Since the x's are homogeneous, any point of the affine space $x_5 \neq 0$ may be expressed in the form $(x_1, x_2, x_3, x_4, 1)$; then the first four x's serve as nonhomogeneous affine coordinates and can be identified with the familiar x, y, z, ct of the special theory of relativity. The light cone at any event (a) is

$$(x_1 - a_1)^2 + (x_2 - a_2)^2 + (x_3 - a_3)^2 - (x_4 - a_4)^2 = 0.$$

The line joining two events, $A = (a)$ and $B = (b)$, is timelike, null, or spacelike according as the number

$$(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 - (a_4 - b_4)^2$$

is negative, zero, or positive; and the interval or *separation* AB between these two events is defined to be the square root of the absolute magnitude of this number.

For comparison with Synge [57, p. 40], note that instead of (x_1, x_2, x_3, x_4) we should strictly write (x^1, x^2, x^3, x^4). In fact,

Synge uses (x_1, x_2, x_3, x_4) for the so-called Minkowskian coordinates, which do not concern us at all [57, p. 56].

As a diagram for relativistic cosmology in the large, Minkowski's world has been almost entirely superseded by de Sitter's, which, as Synge says, "opens up new vistas, introducing us to the idea that space (a slice of space-time) may be finite, and this seems to satisfy some mental need in us, for infinity is one of those things which we find difficulty in comprehending" [58, p. 257]. Timelike lines remain infinite, but most of us are willing to accept eternity.

In the same year that Eddington published the second edition of *The Mathematical Theory of Relativity* [28], Du Val [27] made a discovery that has been almost completely ignored, even by Robb and Synge. This discovery is that de Sitter's world, in its *polar* or *elliptic* form [58, p. 260] is exterior-hyperbolic four-space: the part of real projective four-space that lies outside a nonruled quadric hypersurface.

At any event A, the light cone envelops the absolute quadric; in other words, the null lines are tangents, the timelike lines are secants, and the spacelike lines are nonsecants (Figure 7). Study's observation of the "nontriangle inequality" for a triangle of timelike lines is thus revealed as an early version of Einstein's *twin paradox:* An astronaut flies away, at a speed $c - \epsilon$, to some distant planet, and returns, still in the prime of life, to find that his twin brother, who stayed at home, is now an old man.

The polar hyperplane α of A intersects the quadric hypersurface in a quadric surface, and allows us to represent the lines of all three

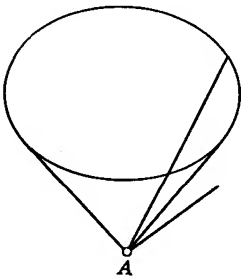


Figure 7

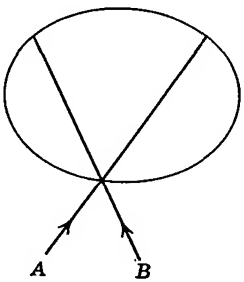


Figure 8

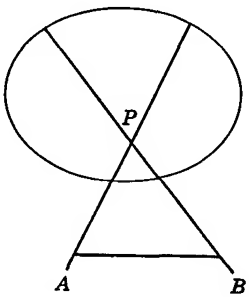


Figure 9

kinds through A by the points that are their α-sections, just like the hyperplane at infinity in Minkowski's world. It is only when we pass from A to another "observer" B that differences are seen: the null lines through B are no longer "parallel" to those through A. But the adjective *parallel* can still be used for two timelike lines that have the same point at infinity, that is, two lines that meet on the quadric hypersurface (as in Figure 8), so that their extensions inside the quadric are parallel in the sense of classical hyperbolic geometry. If such lines are the world lines of two unaccelerated particles A and B, the spacelike interval AB does not remain constant but decreases asymptotically. On the other hand, two secants which have in common an interior point P (as in Figure 9) are the world lines of unaccelerated particles which have a minimax interval on the polar hyperplane of P and afterwards diverge.

This use of the word *minimax* has the following justification. For each position of A on the first world line there is a corresponding position of B on the second, namely the foot of the *perpendicular* from A in the sense of the exterior-hyperbolic geometry. This is the point on the second world line that is *farthest* from A. (Two other positions of B, lying on tangents from A to the quadric, are actually at *zero* distance from A.) This maximum distance AB attains its minimum when A is conjugate to P, so that AB is perpendicular to both the world lines.

Similar remarks apply to *skew* world lines [19] which do not lie in a plane (but, of course, in a hyperplane). Thus two random particles (with an infinitesimal probability of having parallel world lines) will eventually diverge, and "a number of particles initially at rest will tend to scatter" [28, p. 161]. In fact, we can deduce the "expanding universe" from Du Val's projective diagram without invoking any calculation.

When we do wish to make calculations, we naturally take the absolute quadric to be

$$x_1^2 + x_2^2 + x_3^2 - x_4^2 + x_5^2 = 0.$$

Since the x's are homogeneous, we can normalize the coordinates so that every event (x) satisfies

$$x_1^2 + x_2^2 + x_3^2 - x_4^2 + x_5^2 = 1.$$

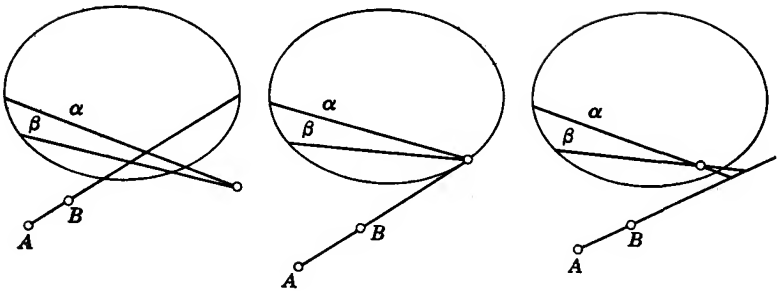


Figure 10

Then the light cone at any event (a) is

$$x_1^2 + x_2^2 + x_3^2 - x_4^2 + x_5^2 - (a_1x_1 + a_2x_2 + a_3x_3 - a_4x_4 + a_5x_5)^2 = 0;$$

for example, the light cone at $(0, 0, 0, 0, 1)$ is

$$x_1^2 + x_2^2 + x_3^2 - x_4^2 = 0,$$

exactly as in the Minkowskian case. The line joining two events, $A = (a)$ and $B = (b)$, is timelike, null, or spacelike according as their polar hyperplanes α and β (in the ordinary or "interior" hyperbolic space) are ultraparallel (having the line AB as their common perpendicular), parallel, or intersecting (see Figure 10), that is [16, p. 193], according as the cross ratio

$$\{AB, \beta\alpha\} = (a_1b_1 + a_2b_2 + a_3b_3 - a_4b_4 + a_5b_5)^2 = (ab)^2$$

is greater than 1, equal to 1, or less than 1; and the separation AB between these two events is given by

$$\cosh AB = |(ab)|, \qquad AB = 0, \qquad \text{or} \qquad \cos AB = |(ab)|,$$

respectively [17, p. 281]. It follows [17, p. 249] that the metric form Φ [58, p. 1] is

$$dx_1^2 + dx_2^2 + dx_3^2 - dx_4^2 + dx_5^2,$$

where dx_4 or dx_5 can be eliminated (if we so desire) by means of the relation

$$x_1\,dx_1 + x_2\,dx_2 + x_3\,dx_3 - x_4\,dx_4 + x_5\,dx_5 = 0.$$

Synge [58, p. 262] described de Sitter's world as a space-time of

constant positive curvature. We have simplified the discussion by choosing such units of measurement that this curvature is 1 and the velocity of light is 1. In other words, our unit of distance is "the radius of curvature of the universe" and our unit of time is the time light would take to travel this distance. It must be admitted that this projective approach yields only the nonorientable "polar" form of de Sitter's world, whereas Synge prefers its orientable covering manifold, the "antipodal" form. However, all the above formulae remain valid; the only difference is that two antipodal events, such as $(0, 0, 0, 0, \pm 1)$, are regarded as being distinct instead of coincident.

CONCLUSION

We have considered projective geometries of two or more dimensions over various fields, and the possibility of replacing the field by a corpus or a ternary ring. We might just as well have considered affine, Euclidean, non-Euclidean, and pseudo-Euclidean geometries over the same fields. Thus there are many geometries, each describing another world: wonderlands and Utopias, refreshingly different from the world we live in.

I hope that these excursions, in a few of many possible directions, have helped to reveal the healthy state of development of this fascinating subject, including its interactions with other branches of pure and applied mathematics.

REFERENCES

1. A. D. Alexandrow, *Die innere Geometrie der konvexen Flächen*, Berlin (1955) Akademie-Verlag.
2. Emil Artin, *Geometric algebra*, New York (1957) Interscience.
3. Friedrich Bachmann, *Aufbau der Geometrie aus dem Spiegelungsbegriff*, Berlin (1959) Springer.
4. Reinhold Baer, Polarities in finite projective planes, *Bull. Amer. Math. Soc.*, **52** (1946) 77–93.
5. W. W. Rouse Ball, *Mathematical recreations and essays*, London (1959) Macmillan.
6. A. S. Besicovitch, The Kakeya problem, *Amer. Math. Monthly*, **70** (1963) 697–706.
7. H. F. Blichfeldt, The minimum values of quadratic forms in six, seven and eight variables, *Math. Z.*, **39** (1935) 1–15.
8. F. van der Blij, History of the octaves, *Simon Stevin*, **34** (1961) 106–125.

8a. Max Brückner, Ueber die Ableitung der allgemeinen Polytope und die nach Isomorphismus verschiedenen Typen der allgemeinen Achtzelle (Oktatope), *Verk. Nederl. Akad. Wetensch.*, Sect. I, 10.1 (1909).
9. R. H. Bruck and E. Kleinfeld, The structure of alternative division rings, *Proc. Amer. Math. Soc.*, **2** (1951) 878–890.
10. R. H. Bruck and H. J. Ryser, The non-existence of certain finite projective planes, *Canad. J. Math.*, **1** (1949) 88–93.
10a. C. Carathéodory, Ueber den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen, *Rend. Circ. Mat. Palermo*, **32** (1911) 193–217.
11. Claude Chevalley, Démonstration d'une hypothèsis de M. Artin, *Abh. Math. Sem. Hamburg*, **11** (1936) 73–75.
12. H. S. M. Coxeter, Integral Cayley numbers, *Duke Math. J.*, **13** (1946) 561–578.
13. H. S. M. Coxeter, Review of a paper by Melmore, *Math. Rev.*, **9** (1948) 53.
14. H. S. M. Coxeter, Self-dual configurations and regular graphs, *Bull. Amer. Math. Soc.*, **56** (1950) 413–455.
15. H. S. M. Coxeter, Extreme forms, *Canad. J. Math.*, **3** (1951) 391–441.
16. H. S. M. Coxeter, *The real projective plane*, Cambridge, England (1955) Cambridge University Press.
17. H. S. M. Coxeter, *Non-Euclidean geometry*, Toronto (1961) University of Toronto Press.
18. H. S. M. Coxeter, Twelve points in PG(5, 3) with 95040 self-transformations, *Proc. Royal Soc.*, A **247** (1958) 279–293.
19. H. S. M. Coxeter, On Wigner's problem of reflected light signals in de Sitter space, *Proc. Royal Soc.*, A **261** (1961) 435–442.
20. H. S. M. Coxeter, *Introduction to geometry*, New York (1961) Wiley.
21. H. S. M. Coxeter, An upper bound for the number of equal nonoverlapping spheres that can touch another of the same size, *Proc. Sympos. Pure Math.*, **7** (1963) 53–72 (Amer. Math. Soc., Providence, R.I.).
22. H. S. M. Coxeter, *Regular polytopes*, New York (1963) Macmillan.
23. H. S. M. Coxeter, *Projective geometry*, New York (1964) Blaisdell.
24. H. S. M. Coxeter and W. O. J. Moser, *Generators and relations for discrete groups*, Berlin (1965) Springer's Ergeb. Math. N.F. 14.
25. Ludwig Danzer, Branko Grünbaum and Victor Klee, Helly's theorem and its relatives, *Proc. Sympos. Pure Math.*, **7** (1963) 101–180 (Amer. Math. Soc., Providence, R.I.).
26. L. E. Dickson, *Linear groups, with an exposition of the Galois field theory*, Leipzig (1901).
27. Patrick Du Val, Geometrical note on de Sitter's world, *Phil. Mag.* (6) **47** (1924) 930–938.
27a. Patrick Du Val, *Homographies, quaternions, and rotations*, Oxford, England (1964) Clarendon Press.
28. A. S. Eddington, *The mathematical theory of relativity*, Cambridge, England (1924) Cambridge University Press.
29. Gino Fano, Sui postulati fondamentali della geometria proiettiva in uno spazio lineare a un numero qualunque di dimensioni, *Giorn. Mat.*, **30** (1892) 106–132.

30. László Fejes Tóth, *Regular figures*, Oxford, England (1964) Pergamon.
31. Hans Freudenthal, Zur ebenen Oktavengeometrie, *Proc. Kon. Akad. Wet. Amsterdam*, A **56** (1953) 195–200.
32. David Gale, Neighborly and cyclic polytopes, *Proc. Sympos. Pure Math.*, **7** (1963) 225–232 (Amer. Math. Soc., Providence, R.I.).
33. C. F. Gauss, *Werke*, Göttingen (1876).
34. A. Gilbert and D. MacDonnell, The Steiner-Lehmus theorem, *Amer. Math. Monthly*, **70** (1963) 79–80.
35. A. M. Gleason, Finite Fano planes, *Amer. J. Math.*, **78** (1956) 797–807.
35a. Hugo Hadwiger, Ungelöste Probleme, Nr. 17, *Elemente der Math.*, **12** (1957) 61–62.
36. Marshall Hall, Jr., *The theory of groups*, New York (1959) Macmillan.
37. Marshall Hall, Jr., J. D. Swift, and R. J. Walker, Uniqueness of the projective geometry of order eight, *Math. Tables Aids Comput.*, **10** (1956) 186–194.
38. P. J. Heawood, On the four-colour map theorem, *Quarterly J. Math.*, **29** (1897) 270–285.
39. L. Heffter and C. Koehler, *Lehrbuch der analytischen Geometrie* I, Karlsruhe (1927).
40. Irving Kaplansky, Lie algebras, in Saaty (Ed.), *Lectures on Modern Mathematics* I, New York (1963) Wiley, 115–132.
41. Sidney Melmore, Densest packing of equal spheres, *Nature*, **159** (1947) 817.
42. Hermann Minkowski, Diskontinuitätsbereiche für arithmetische Äquivalenz, *J. Reine Angew. Math.*, **129** (1905) 220–274.
42a. Hermann Minkowski, Raum und Zeit, in H. A. Lorentz, A. Einstein, and H. Minkowski, *Das Relativitätsprinzip*, Leipzig (1922) Teubner, pp. 54–71.
43. F. R. Moulton, A simple non-Desarguesian plane geometry, *Trans. Amer. Math. Soc.*, **3** (1902) 192–195.
44. Hanna Neumann, On some finite non-Desarguesian planes, *Archiv Math.*, **6** (1954) 36–40.
45. T. G. Ostrom, Separation, betweenness, and congruence in planes over non-ordered fields, *Archiv Math.*, **10** (1959) 200–205.
46. Daniel Pedoe, *An introduction to projective geometry*, New York (1963) Macmillan.
47. Günter Pickert, *Projektive Ebenen*, Berlin (1955) Springer.
48. A. A. Robb, *Geometry of time and space*, Cambridge, England (1936) Cambridge University Press.
48a. C. A. Rogers, *Packing and covering*, Cambridge, England (1964) Cambridge University Press.
49. Bertrand Russell, *The principles of mathematics*, London (1937).
50. Ludwig Schläfli, *Gesammelte mathematische Abhandlungen* I, Basel (1950) Birkhäuser.
51. Beniamino Segre, Plans graphiques algébriques réels non désarguesiens et correspondances crémoniennes topologiques, *Rev. Math. pures appl.*, **1.3** (1956) 35–50.
52. Beniamino Segre, Le geometrie di Galois, *Ann. di Mat.*, (4) **48** (1959) 1–96.

53. Beniamino Segre, *Lectures on modern geometry*, Rome (1961) Cremonese.
54. Robert Steinberg, Finite reflection groups, *Trans. Amer. Math. Soc.*, **91** (1959) 493–504.
55. Eduard Study, Beiträge zur nicht-euklidische Geometrie, *Amer. J. Math.*, **29** (1907) 101–167.
56. J. L. Synge, An introduction to space-time, *New Scientist*, May 1, 1958.
57. J. L. Synge, *Relativity: The special theory*, Amsterdam (1956) North-Holland.
58. J. L. Synge, *Relativity: The general theory*, Amsterdam (1960) North-Holland.
59. G. Tarry, Le problème des 36 officiers, *C.R. Assoc. Fr. Av. Sci.*, **2** (1901) 170–203.
60. J. A. Todd, On representations of the Mathieu groups as collineation groups, *J. London Math. Soc.*, **34** (1959) 406–416.
61. Oswald Veblen and W. H. Bussey, Finite projective geometries, *Trans. Amer. Math. Soc.*, **7** (1906) 241–259.
62. A. N. Whitehead, *The axioms of descriptive geometry*, Cambridge, England (1907) Cambridge University Press.
63. Ernst Witt, Über die Kommutativität endlicher Schiefkörper, *Abh. Math. Sem. Hamburg*, **8** (1931).
64. Ernst Witt, Spiegelungsgruppen und Aufzählung halbeinfacher Liescher Ringe, *Abh. Math. Sem. Hamburg*, **14** (1941) 289–322.
65. J. W. Young, *Monographs on topics of modern mathematics*, New York (1911) Longmans Green.

3

Mathematical Logic

Georg Kreisel

The bulk of the material in this chapter consists of technical mathematical results, some of which have pure mathematical interest. But they are primarily selected for their use in logical or foundational studies, for which the subject of mathematical logic was originally introduced.

The data of foundations consist of the mathematical experience of the working mathematician. This presents itself as being about certain mathematical *objects* such as natural numbers or groups; the processes involved in mathematical activity present themselves as primarily *mental* and *abstract* (not sensory), but they are recorded and communicated by use of *concrete*, that is, spatiotemporal, *symbols*. The three elements of mathematical experience just mentioned are familiar from the so-called *realist*, *idealist*, and *formalist* views in the philosophy of mathematics, though, traditionally, the latter are presented as *conflicting* views on the meaning of mathematical assertions: cf. below.

The general problem of foundations is to analyze this experience taken as a *whole* or, at least, large portions of it. The following questions are regarded as basic for such an analysis: which, if any,

of the three elements of mathematical experience above are primary and which are subsidiary? Depending on the answer: what grammatical rules lead to meaningful expressions? and: what principles of proof or, more generally, of evidence are valid?[1] These problems and several distinctions that are basic in an analysis of the *whole* of mathematics have so far not been found to be equally important in the analysis or "foundation" of special branches such as familiar number theory or topology. As the subject progresses new important problems can be formulated.

Two kinds of reactions to these questions are familiar. One regards it as hopeless to find an adequate and precise formulation of foundational problems; perhaps because some questions which seemed quite precise turned out to be ambiguous in *some* respects. The other is satisfied with quite superficial answers such as old-fashioned formalism (cf. Section 3 below); perhaps because the questions, though old and therefore natural, have progressed little for a long time.

The main theme of the present exposition is to make out a case that real, if modest, progress has been made on foundational problems. It is not a mere one-shot success story: at least some of the informal problems and conceptions did not collapse when unexpected ambiguities and other difficulties were discovered. They became stable once a few basic distinctions were made. Moreover, generally the revisions needed were essentially unambiguous. Certainly, no less cogent than the steps which have led, for example, from the atomic conception of matter at the time of the Greeks to its modern formulation. The particular conception which we use most to illustrate this is the well-known

Formalist or Mechanist Conception of Mathematics. This conception contends that, though reasoning does not present itself to us as a manipulation of symbols—for example, words—according to mechanical rules, but rather as a series of insights about some imagined concrete or abstract structures, all reasoning is fully

[1] And, in the case of mathematics, these questions and their answers do not exhaust the questions about reasoning that present themselves to us naturally, that is, on the phenomenalistic level, particularly those concerning *intelligibility*. Here one does not primarily ask what mathematics is *about*, but tries to make the structure of complicated sets of mathematical facts graspable, as, for example, in the algebraic approach to (parts of) analysis. In a logical analysis, concerned with the validity of reasoning, one must ask *what* is asserted.

reflected in the manipulation of symbols and their combinatorial properties. This contention, of course, is consistent with very general positivistic views; but in the domain of mathematics, it is supported by a remarkable *discovery*, not at all obvious from a superficial examination of ordinary mathematical practice: formalization, that is, the existence of simple mechanical rules (of predicate logic and set theory) from which the formalized versions of ordinary theorems can be derived. Of course, the discovery itself was not formal or linguistic, since the theorems of ordinary language and their formal representations do not resemble each other in external (syntactic) form, but in their meaning.[2] Formalism does not deny this, but only claims that the *significant questions about reasoning can be treated as combinatorial questions about the mechanical rules,*[3] *and solved by combinatorially evident methods;* and thus all abstract existential assumptions are eliminated. For details, cf. Section 3.

Irresoluble Conflicts. The first kind of defeatist mentioned above foresees hopeless ambiguity in what would be regarded as significant, and perhaps even in the meanings of "combinatorial" or "mechanical"; the other rejects all unanswered questions as insignificant! Therefore, for both of them the problem of the correctness of the formalist contention is not real. But although the theoretical possibility of conflict exists, *experience of actual research* shows that in the present case conflicts have been satisfactorily solved.

On the one hand, Hilbert's program, described in Section 3, shows how to formulate (for a suitably sophisticated formalist conception) questions of *evidence*, which, at first sight, seems quite difficult. And, further, for an astonishingly wide area of mathematics (3.3232) they are solved positively, that is, Hilbert's program can be carried out. On the other, for negative results, it is necessary, of course, to give a searching analysis of the meanings of "combinatorial" and

[2] Indeed, the mechanical rules were not found inductively, but by formulating properties of the abstract concepts, like *set* or *construction*. This is one reason why we put the abstract theory of these notions first (Sections 1 and 2), and only afterwards (Section 3) the formalistic (proof theoretic) study of the axioms so found.

[3] Somewhat like the aim of elementary mechanics to represent planets as point masses, ignoring their physical composition and even their shape; it is known that the finer details of even their paths cannot be got in this way.

"mechanical": for the analyses of 3.4 and 3.5 (involving in a modest way the questions of footnote 1), it is shown, in conjunction with Gödel's incompleteness theorem, that Hilbert's program cannot be carried out for certain specific parts of current mathematics. The simple-minded mechanistic position is refuted by the incompleteness theorem itself (3.2341).

It should be stressed here that mathematical logic is essential in foundational problems not only in the advanced development, but in the precise formulation of logical theories; just as the discipline of partial differential equations is needed to formulate modern physical theories.

Other Traditional Views. Particularly in view of the limitations of the mechanist conception, Sections 1 and 2 do not only serve the purposes mentioned in footnote 2, but are a contribution to other traditional conceptions of mathematics, namely so-called realism (or platonism, according to which also *ideas* are real, that is, external to ourselves) and idealism (or mentalism, according to which mathematical objects are thoughts). Here the situation is much less satisfactory than in the case of formalism; in particular, the known precise developments, namely those of Section 1 and 2, are remarkably *special* versions of realism and idealism respectively. Also realism and idealism are truly in opposition only if taken in the *strict* sense, as all-embracing conceptions, that is, that mathematics is *only* about objects external, not external to us. But they do not seem to be sufficiently developed to decide between them (or against both) nearly as well as in the case of formalism. We postpone the discussion to the end (Section 5) where we can refer to the technical machinery and specific results of the text, and where general open questions are formulated.

SET THEORY

The reader will be familiar with the standard *techniques* of set theory. He will find a beautifully readable discussion of the basic notions in [35], and detailed treatments in [1; 9]; in particular, a full development of the particular laws about sets known as Zermelo's, respectively Zermelo-Fränkel's set theory.

In any deductive exposition, one's view of the nature of the subject enters at the beginning; the rest is development. The realist

view of mathematics, in particular of its objectivity, determines the choice of (syntactic) rules for forming *meaningful* expressions and for making *valid* inferences. Implicit in the requirement of objectivity is this: symbols are to denote *well-defined* objects $a, b \ldots$ and relations R, and so $R(a, b)$ is either *true* or *false*; also, if c is a *well-defined collection* and $(\forall b \in c)$ stands for: for all elements b of c, then $(\forall b \in c)R(a, b)$ is a well-defined property (of a). Similarly for the other logical operations: negation $\bar{}$, conjunction $\cdot$, implication $\Rightarrow$, existential quantification $\exists$ [note: $(\forall x \in c)A(x) \Leftrightarrow \overline{(\exists x \in c)\overline{A(x)}}$]. Part of the "operational significance" of the requirement of objectivity is that the laws of two-valued (or classical) logic hold.

The so-called set theoretic foundation of mathematics specializes this general requirement in two opposite directions. It asserts the existence of (well-defined) *infinite* collections, and it restricts itself to a *single* basic relation (membership: $\in$). This restriction not only does not follow from the realist view, but is *a priori* not even plausible. It is supported by researches of the last century (reduction of natural numbers and real numbers to set theory) which the reader is assumed to know. It is to be remarked that the reduction was carried out for these two *concrete* notions, and not for *abstract structures* (cf. 1.8 below).

As stated above, the use of classical logic is implicit in the realist view, and so are the laws about sets which are used in current mathematical *practice*. The converse cannot be expected to hold, that is, that the view is implicit in the particular laws used in practice. (To consider alternative accounts of this practice is in fact the object of Hilbert's program, cf. Section 3.) To what extent the laws used express not only objectivity, but that mathematical objects, that is, the particular objects considered here, are *external* to us, is discussed in Section 5. At the present stage we confine the analysis of the basic notions to the following remarks on

Collections, Properties, or Abstract Objects (Structures)? Our intention here is to formulate laws that are valid when the variables denote collections of objects which are themselves collections, and "$\in$" denotes membership. We do not require that the laws also hold if the variables are interpreted to range over properties and $a \in b$ is read as: a has the property b, or: the variables range over abstract structures and $a \in b$ is read as: the structure a is an

instance of the structure b (for example, a = the structure of the integers with addition and multiplication, b = general ring structure). Even if these two interpretations were sufficiently clear to make such a common theory feasible, we could not expect to get as many valid laws as for just one interpretation. The only question is: is the notion of collection not dependent on these other notions? For example, as is sometimes supposed, is it a mere analogue to finite collection, or an abstraction from conditions or properties? One speaks of a wood (a collection) without or before having counted the trees; so, finiteness is not essential. In fact, generally, the *obvious* "property" satisfied by just the elements of a collection C is simply: belonging to C. Next, for a collection to be regarded as an abstraction from (or the extension of) a property P, P either has to be defined for *arbitrary objects* or else *limited* in advance to some collection C'.

In the latter case, C' has to be presupposed, and this is indeed the usual case in mathematics, where one considers sets *of* something, for example, of numbers or of sets of numbers. Here the primary step is to formulate basic principles of *generating* collections (below: collecting into one collection $\mathfrak{P}(x)$ the subcollections of any given x); properties on particular C' can then be explicitly defined in terms of the collections so generated. This is the principle of the *cumulative type theory* below, except that, for a smooth development, one takes the process: $x \to x \cup \mathfrak{P}(x)$ instead. Zermelo's axioms are valid if one starts with an infinite collection of individuals, and iterates this construction ω times. As emphasized in [34] the restriction on the comprehension axiom is *not* ad hoc; in *this* structure, in general there simply is no collection x consisting of precisely those collections which satisfy a given condition Φ; but by the subset construction there is one satisfying $\Phi \cdot x \in a$ if a is in the structure. (Of course, $(\exists y)\forall x[x \in y \Leftrightarrow \Phi(x)] \Rightarrow (\exists a)(\forall x)[\Phi(x) \Rightarrow x \in a]$.)

In the former case there is a genuine problem: what properties define collections, particularly if properties themselves are to be regarded as objects. This may be shown by means of the *paradoxes*. If x is the collection of objects satisfying a property, take $\{x\}$ (whose only element is x) to be the property regarded as an object. Let r (for Russell) be *any* collection satisfying $(\forall x)(x \in r \Rightarrow x \notin x)$, and so $r \notin r$. Thus, $\forall x(x \in r \cup \{r\} \Rightarrow x \notin x)$, $r \cup \{r\} \supset r$ but $r \cup \{r\} \neq r$. In other words, r is not the collection

of all objects satisfying $x \notin x$.[4] This result has been (properly) used to refute a too simple-minded (platonist) realist view.

1.1. *Cumulative Theory of Types.* Zermelo's informal derivation of his axioms can be analyzed by formulating explicitly properties of the cumulative-type structure described above. The analysis is in [54], not only for these axioms, but for several other important cases. The basic notions in terms of which the construction is described, are:

Set (collection); $a, b, \ldots ; x, y, \ldots$ (are the corresponding variables)

Iterations (types); $\alpha, \beta, \ldots ; \xi, \eta, \ldots$

Equality: $x = y$, $\alpha = \beta$; membership: $x \in y$; (strict) order: $\xi < \eta$; x is of type ξ, denoted by $(x : \xi)$. Note: "type" means cumulative type, that is, x is obtained by $\leq \xi$, not only exactly ξ iterations.

Basic laws (NB. They are properties of the structure, not intended to "define" it axiomatically!) Unlimited quantifiers $\forall \xi$, $\exists \xi$ are intended to range over all ordinals $<$ some given one α_0, $\forall x$, $\exists y$ over sets of the type structure of type $< \alpha_0$. Otherwise—for a set theoretic foundation in the strict sense, cf. page 98—the meaning of quantified expressions would be well defined only on the assumption that there is a collection consisting of all collections, which is false for the intended structure. It is not assumed that sets and ordinals are necessarily distinct kinds of objects. For individuals a, we use the convention $\forall x(x \in a \Leftrightarrow x = a)$ to ensure extensionality. For all a, b, α, β (considered) and *formulae* Φ

$\forall x(x \in a \Leftrightarrow x \in b) \Rightarrow a = b$ (extensionality)

$\exists y \forall x\{x \in y \Leftrightarrow [(x : \alpha) \cdot \Phi(x)]\}$ for all Φ not containing y (comprehension), though arbitrary properties Φ are *intended;*

$\alpha < \beta \Leftrightarrow \overline{(\alpha = \beta \cdot \beta < \alpha)}$ (trichotomy), $(\exists \xi)(\alpha < \xi)$,

$\exists \xi \Phi(\xi) \Rightarrow \exists \xi\{\Phi(\xi) \cdot \overline{\exists \eta[\eta < \xi \cdot \Phi(\eta)]}\}$ for all Φ (least element)

$\forall x\{(x : \alpha) \Leftrightarrow \exists \xi(\xi < \alpha \cdot \forall y[y \in x \Rightarrow (y : \xi)])\}$ (collection)

$\exists \xi(a : \xi)$, $\exists x \forall y[(y : \alpha) \Leftrightarrow y \in x]$ (type structure)

Infinite types are ensured by:

$$\exists \omega \exists \xi \forall \eta \exists \zeta\{\xi < \omega \cdot [\eta < \omega \Rightarrow (\eta < \zeta \cdot \zeta < \omega)]\}.$$

[4] One of the set theoretic *definitions* of ordinals takes the empty set $\emptyset$ to be zero: $a \to a \cup \{a\}$ as the successor function (and unions for limits). The argument above is literally the proof that there is no greatest natural number (greatest ordinal).

A more recondite, but quite basic, principle asserts the existence of a *supremum.*

$$(\forall \xi < \alpha)\exists \eta \Phi(\xi, \eta) \Rightarrow (\exists \zeta)(\forall \xi < \alpha)(\exists \eta < \zeta)\Phi(\xi, \eta) \qquad \text{for all } \Phi.$$

This "ordinal theoretic" principle is equivalent (by use of the remaining axioms) to the *reflection principle* $\forall \xi \exists \eta(\xi < \eta \cdot (\forall x{:}\eta)(\forall y{:}\eta) \ldots \{\Phi(x, y, \ldots) \Leftrightarrow \Phi^{(\eta)}(x, y, \ldots)\})$ where $\Phi^{(\eta)}$ is obtained from Φ by restricting all variables in Φ to be of type η: what is true in the whole universe under consideration is already true at some $\eta(\eta < \alpha_0)$.

The proof of the reflection principle from the supremum contains what might be described as *the* method of the subject! It is fully illustrated by considering purely existential Φ, say $\exists z \psi(x, z)$, ψ quantifier free: for each x of type ξ, either $\exists z \psi(x, z)$ holds in the structure or there is a least type η_x for which: $\exists z[(z{:}\eta_x) \cdot \psi(x, z)]$. Take the supremum ξ_1, and iterate. Take for η the first fixed point of $\xi \rightarrow \xi_1$.

1.2. *First Main Result.* The logical consequences of (the properties of the cumulative type structure in) 1.1. which are purely set theoretic, that is, those expressed in terms of $\in$ and lower-case letters, are precisely those of Zermelo-Frankel's set theory [53].

Similarly one obtains an axiomatization of all ordinal theoretic consequences, that is, those expressible in terms of $<$ and lower-case Greek letters.

1.3. *Converse Results* (reductions to set theory). Naturally one cannot expect to prove that the ordinals *are* sets or vice versa! But the familiar *definitions* of ordinals in set theory show:

1.31. There are purely set theoretic conditions $O(x)$, $L(x, y)$, $T(x, y)$ [x is an ordinal, x precedes y, x is of type y and $O(y)$, respectively] such that the laws of 1.1 are formal *theorems* of set theory when Greek variables are restricted to range over O, $\alpha < \beta$ is replaced by $L(\alpha, \beta)$, and $(x{:}\alpha)$ by $T(x, \alpha)$.

1.32. (Strengthening). Dedekind showed that if a, $b(b \subset a^2)$, a', b' are two pairs of sets which satisfy Peano's axioms (a is the domain of the order relation b), then a, a' are isomorphic so as to map: $b \rightarrow b'$. In other words, Peano's axioms determine the structure uniquely. Without the axiom of infinity, one can restate the the result as follows: If $N(x)$, $L(x, y)$ and $N'(x)$, $L'(x, y)$ are two pairs of conditions for which Peano's axioms can be *proved*, then an

isomorphism can be defined which maps N onto N', preserving order. Presumably the same holds for O, L above.

The other converse result to be considered is this: to give purely ordinal theoretic conditions $M(\alpha)$, $E(\alpha, \beta)$, $T_1(\alpha, \beta)$(α is a set, $\alpha \in \beta$, the set α is of type β respectively) such that the axioms of 1.1 are satisfied under this interpretation. This is a more delicate matter, cf. 1.52; also, constants for certain ordinal functions have to be introduced since by use of $<$ only ordinals less than ω^ω can be defined.

1.33. There are several quite interesting refinements of these reductions in the literature relating natural fragments of set theory to natural schemes of axioms for other familiar mathematical notions; for example, first order arithmetic (Z of [6]) and set theory without the axiom of infinity; or classical analysis (Suppl. IV of [6]) and set theory with the axiom of infinity, but without the power set axiom.

1.4. *Changing the Parameters in 1.1.* There are three things that are naturally changed: the *starting* collection C_0 of individuals (possibly empty), the *number* α_0 of iterations, the kinds of *subcollection* taken at each stage.

1.41. Evidently the axioms 1.1 hold for an arbitrary C_0, all α_0 satisfying suitable *closure* conditions, for example, α_0 any limit number $>\omega$ for Zermelo's own theory, and taking *all* subcollections. This was simply the intention.

1.42. (Almost evidently) since we have only denumerably many axioms in 1.1, by a simple application of the Skolem-Loewenheim argument, even the supremum axiom is satisfied for a suitable α_0 by taking suitable denumerable subcollections at each $\alpha < \alpha_0(\bar{\bar{\alpha}}_0 = \aleph_0)$.

The problem one sets oneself is this: to find *manageable* restrictions on the collecting process for which the axioms hold; or, more specifically, restrictions which yield *minimal* models. Compare the case of Euclidean space. The intended structure consists of *all* points (or, in terms of coordinates, of *all* triples of real numbers). However, for the study of particular laws, that is, of a particular axiomatic theory, it is necessary to introduce manageably restricted models, for example, Pythagorean fields, for showing the impossibility of ruler-compass solutions.

Useful specializations of parameters in 1.1, which will be used below, are these.

1.431. $C_0 = \emptyset$: all sets are well founded, that is, there are no infinite descending $\in$ chains.[5]

1.432. $C_0 = \emptyset$, $\alpha_0 = \omega$. These are the so-called *hereditarily finite* sets. They satisfy all the axioms except of course the axiom of infinity or the reflection principle; the replacement axiom expresses that a function with a finite range has a finite domain.

1.433. $C_0 = \emptyset$, $\alpha_0 = \Omega$ (first uncountable ordinal), take *all countable* subcollections at each stage. Now the power set axiom is not satisfied, but the replacement axiom is, the countable axiom of choice being satisfied for the intended structure (cf. [35]).

1.5. *Second Main Result.* Let the *ramified hierarchy on* C_0 be defined as follows, for ordinals α: (each R_α, $\alpha > 0$, is a collection of collections)

$R_0 = C_0$,

For limit numbers α, $R_\alpha = \bigcup_{\beta<\alpha} R_\beta$

$R_{\alpha+1}$: the union of R_α and of all subsets of R_α explicitly definable by means of symbols for particular elements of R_α, $a \in b$, propositional connectives and quantifiers over R_α. For the connection with Russell's theory of orders (Poincaré's predicativity) cf. [33].

All familiar axiom schemes for set theory are satisfied by suitable R_α. More precisely, for $C_0 = \emptyset$,

1.511. Zermelo's axioms are satisfied at $\aleph_\omega$

1.512. Zermelo-Fränkel's axioms are satisfied at the first inaccessible[6] ordinal $>\omega$, that is, the first ξ at which the full cumulative type structure satisfies them (in their intended form, cf. page 101).

1.513. The results extend to additions of axioms expressing the existence of larger inaccessible ordinals.

The proof [5] consists of formulating (set theoretically) the definition of the sequence R_α by transfinite induction, that is, giving

[5] In this case the axiom of regularity (Fundierung) is valid.

[6] In traditional terminology; the least ξ such that for all η, ζ ($\bar{\bar{\eta}} < \bar{\bar{\xi}}, \bar{\bar{\zeta}} < \bar{\bar{\xi}}$) and all sequences $\alpha_\rho (\bar{\bar{\alpha}}_\rho < \bar{\bar{\xi}})$: $\left(\sum_{\rho<\eta} \bar{\bar{\alpha}}_\rho\right) < \bar{\bar{\xi}}$, $\bar{\bar{\eta}}^{\bar{\bar{\zeta}}} < \bar{\bar{\xi}}$. ω is inaccessible, so is 2. The terminology is intended to suggest that ξ is critical like ω: The results of 1.7 cast doubt on the reasonableness of this terminology. Why should closure of ξ for these two piddling operations of sums and powers constitute inaccessibility? Does it really express a *significant* property of ω? Note that ω is also the first ordinal with the property: $\forall\alpha(\bar{\bar{\alpha}} < \omega \Rightarrow \bar{\bar{\alpha}} + 1 < \omega)$.

a condition $L(x):\exists\alpha(x \in R_\alpha)$. The proof of 1.511 is slightly different from 1.512 since one cannot use the ordinals of footnote 4, only such ordinals $<\omega+\omega$ being assured by Zermelo's axioms. One uses well orderings instead. Then the proof that the sets satisfying $L(x)$ satisfy the set theoretic axioms is straightforward, by essential use of the argument at the end of 1.1 (and, again, slightly more complicated in 1.511). An *explicit well ordering* of the sets in $L(x)$ is defined by letting x precede y if the first definition of x precedes the first definition of y in the ramified hierarchy. The *generalized continuum hypothesis* ($2^{\aleph_\alpha} = \aleph_{\alpha+1}$) holds when relativized to L, and so do several famous hypotheses from the theory of sets of points. These are the celebrated results of [5].

1.52. *Corollary.* The ordinal theoretic analogue to 1.31 holds, that is, the cumulative theory of types can be interpreted in a theory of ordinals, cf. [70]. The required relation $E(\alpha, \beta)$ for the interpretation of $\in$ is taken to be: the set defined by the αth definition in the ramified hierarchy is an *element* of the set defined by the βth definition, and *this* occurrence of elementhood is expressible by purely ordinal theoretic notions of [71].

In the early literature on set theory there was considerable (and perfectly sensible) interest in the question: is the notion of set definable by (iterated) explicit definitions, as in the ramified hierarchy, more elementary than the principles of set generation in the cumulative hierarchy? The result above clarifies the situation completely as far as familiar principles of set generation are concerned: If only the principles of definition, not the ordinals (iterations) are restricted, it makes no difference from the point of view of consistency.[7] The problem is shifted to questions about the existence of ordinals with certain closure properties. It is not reasonable to use the ramified hierarchy to make the axioms of set theory *itself* more evident. But what it does with wonderful efficiency is to reduce extensions of set theory *to* set theory.

1.53. *Refinements.* There are several significant hierarchies in logic [43]. There is little doubt that all of them are more elegantly presented as segments R_α, in particular as hierarchies of sets of

[7] As noted in [5], the argument described above in the form: the ordinals of a model of set theory S allow one to define a ramified hierarchy satisfying S, can be replaced by a statement with weaker premise and weaker conclusion: the (formal) consistency of S implies quite finitistically the formal consistency of S *together* with the assertion: $\forall x L(x)$.

higher type and not just of sets of natural numbers. More isolated applications are these:

1.531. $C_0 = \{a_0, a_1, \ldots\}$, a collection of individuals without any structure. One easily verifies that all sets of R_α built on such a C_0 are invariant under enough permutations of C_0 to show that the axiom of choice does not hold. (Independence of the axiom of choice when regularity is not assumed.)

1.532. From 1.511, 1.512. Considering the first ordinal $\alpha_Z (\alpha_{ZF}$ respectively) at which $R_{\alpha_{Z(F)}}$ satisfies Zermelo(-Fränkel)'s axioms, one verifies countability[8] of $\alpha_{Z(F)}$ and further that there is a *minimal model* among all models of sets, that is, models in which $\in$ is well founded. For a direct proof, cf. [21].

1.533. Just as in 1.532 Zermelo's axioms give a simple characterisation of α_Z (in fact, the only one we have), so there are simple sets of axioms of set theory which characterize ω_1: the first non-recursive ordinal. Set theory without the power set axiom and with the replacement axiom predicatively restricted is first satisfied at R_{ω_1}. A more significant property [65]: For every $\alpha < \omega_1$, there is an ordering of ordinal α in $R_{\omega+1}$, and all well orderings definable in R_α have ordinal $< \omega_1$.

1.534. It is easily verified that all sets of natural numbers which belong to $R_\alpha (\alpha < \alpha_{ZF})$, are definable in the forms: $\forall S \exists T A(n, S, T)$ *and* $\exists S \forall T B(n, S, T)$, where S, T range over *all* sets of natural numbers, and A, B are purely arithmetic predicates. Conversely, by [63], if a set is so definable it occurs in the (full) hierarchy and furthermore, in this case,

$$(\forall S_{RA})(\exists T_{RA}) A(n, S, T) \Leftrightarrow \forall S \exists T A(n, S, T)$$

where S_{RA}, T_{RA} range only over sets of numbers in the hierarchy: the ranges considered are different, but the sets defined are the same. This leads trivially to the conclusion:

1.535. If a statement $\exists S_{RA} \forall T_{RA} \exists U_{RA} A(S, T, U)$ (A arithmetic) holds, then also $\exists S \forall T \exists U A(S, T, U)$. In proof theoretic form: if it can be proved using the axiom of choice and/or the continuum hypothesis, it can also be proved outright. (Poincaré's

[8] Countable in the full model. While for given C_0 and given abstract ordinal α, R_α is completely determined, a given *description* of an ordinal will in general define different ordinals in different R_α; for example, the description: the first α where R_α satisfies the axioms of Zermelo; no such ordinal exists in R_{α_Z}, but is even countable (by means of mappings) in $R_{\alpha_{ZF}}$.

conjecture is comfortably covered by this!) Thus these axioms are not needed in "concrete" mathematics, though occasionally a proof is *simpler* when they are used, for example in computing homotopy groups.

1.536. Trivially, there are statements $\forall S \exists T \forall U A(S, T, U)$ which are true in the ramified hierarchy but not generally (for example, every set of natural numbers is definable in the ramified hierarchy). It seems to be open whether every *set* $\{n : \forall S \exists T \forall U A(n, S, T, U)\}$ occurs in the ramified hierarchy, of course possibly with a different definition.

The ramified hierarchy uses essentially the *smallest* collecting operation; it has been suggested [35] that the maximal collecting operation, that is, the intended cumulative hierarchy, would yield complementary results, for example, $2^{\aleph_0} > \aleph_1$ (of course, not for the axiom of choice). The situation is asymmetrical because a proof of $2^{\aleph_0} > \aleph_1$ in this structure not only establishes the formal independence of the continuum hypothesis, but *refutes* it. Actually, the formal independence, and much more, was recently established in [22] by a searching analysis of the ramified hierarchy which uses

1.6. *A New Method.* *All* parameters entering into the cumulative type structure are different from the intended structure! The collecting operation is that of the ramified hierarchy, the iteration is stopped at a *countable* α_0 (where R_{α_0} satisfies the axioms of the set theory considered, cf. 1.532), and the starting collection C_0 consists of so-called *generic* sets. Evidently, for $R_{\alpha_0}(C_0)$, that is, the ramified hierarchy starting with C_0, to satisfy $2^{\aleph_0} \neq \aleph_1$ (or simply to be $\neq R_{\alpha_0}$), the sets in C_0 must not be definable in R_{α_0}, though they may be in R_β for some $\beta > \alpha_0$.

The two leading considerations:

1.611. The supremum principle (the basic method of the subject) for ordinals $< \alpha_0$ and relations defined in $R_{\alpha_0}(C_0)$ can be applied to verify the axioms of set theory in $R_{\alpha_0}(C_0)$ only if the part (neighborhood) of C_0 "involved" in any particular axiom of set theory can be defined in R_{α_0}.

1.612. The particular statements (continuum hypothesis, axiom of choice) which are to be false in $R_{\alpha_0}(C_0)$ happen to be *existential*, and so one will try to make it hard to satisfy existential statements.

To simplify the statement of the notions introduced in [22] for handling 1.611, 1.612, recall that just as in 1.52, statements of

$R_{\alpha_0}(C_0)$ can be put into ordinal theoretic form, that is, instead of using quantifiers over the sets defined *from* C_0 (a collection which depends on the unknown C_0), one uses quantifiers over $\alpha < \alpha_0$ (a fixed range).

1.62. *Forcing.* Consider a domain D of individuals and relations on it. Let C be a variable for subsets of D, and let P denote *conditions* on or *neighborhoods* of C [of the form $(d \in C, d' \notin C, \ldots, d'' \in C)$]. Formulae A are built up from primitive formulae and $C(x)$ by means of $\cdot, \bar{}, \exists$. ($\forall A$ is regarded as an *abbreviation* for: $\overline{\exists \bar{A}}$, $A \Rightarrow B$ for: $\overline{A \cdot \bar{B}}$.)

$P \Vdash A$ (read: P *forces* A) is defined by induction:

1.621. A free of C : $P \Vdash A$ if A is true in D.

1.622. $A = C(d)$: $P \Vdash A$ if $\ulcorner d \in C \urcorner$ is in P.

1.623. $A = A_1 \cdot A_2$: $P \Vdash A$ if $P \Vdash A_1$ and $P \Vdash A_2$.

1.624. $A = \bar{B}$: $P \Vdash A$ if, for all consistent extensions P' of P, not $P' \Vdash B$.

1.625. $A = \exists x B(x)$: $P \Vdash A$ if, for some $d \in D$, $P \Vdash B(d)$.

Note that 1.625 is a strong *uniformity* requirement since d depends only on P not on the details of C, cf. 1.612.

Evidently, 1.626: P cannot force both A and $\bar{A}$, 1.627: for every P there is an extension P', $P' \Vdash A$ or $P' \Vdash \bar{A}$, 1.628: by an easy induction, if $P \Vdash A$ and $P' \supset P$, $P' \Vdash A$.

1.63. *Generic Sets.* C_0 is called generic, if for every A in the language considered there is a neighborhood P of C_0 such that $P \Vdash A$ or $P \Vdash \bar{A}$.

1.631. *Existence.* Enumerate the formulae $A_1, A_2, \ldots$; take P_1 so that $P_1 \Vdash A_1$ or $P_1 \Vdash \bar{A}_1$ (1.627), then, $P_2 \supset P_1$, so that $P_2 \Vdash A_2$ or $P_2 \Vdash \bar{A}_2$, and hence (1.628) $P_2 \Vdash A_1$ or $P_2 \Vdash \bar{A}_1$, etc. Take $C_0 = \{d : \exists i (\ulcorner d \in C \urcorner \in P_i)\}$.

1.632. For generic sets C_0, C_0 makes A *true*, if some neighborhood of C_0 *forces* A; in fact, C_0 satisfies A_i if $P_i \Vdash A_i$. (By induction on logical complexity of formulae.) Because of this equivalence between being true and being forced (for generic sets!), by 1.625, we know that a true existential statement satisfies strong uniformity conditions.

Applying this to the case where D consists of the ordinals $< \alpha_0$ and the relations on D are those of 1.52, one satisfies requirement 1.611 by choosing the sequence P_i in some reasonably regular fashion; for example, if $C \subset \omega$ and P are to be *finite* sets, one enumerates the pairs of disjoint finite subsets $\langle s, t \rangle$ of ω, and takes P_1

to be the *first* set of conditions $\{n \in C : n \in s\} \cup \{n \notin C : n \notin t\}$ which $\Vdash A_1$ or $\Vdash \bar{A}_1$. It is easily verified that for any *fixed* A the relation $P \Vdash A$ is definable in R_{α_0}. For each of the usual axioms, the method of supremum shows that $R_{\alpha_0}(C_0)$ satisfies it, if R_{α_0} satisfies (a suitable finite number of) them.

By delicate choice of $C_0 \subset \beta$ (β instead of ω) and use of countability of α_0 (in some R_γ, $\gamma > \alpha_0$), $R_\alpha(C_0)$ can be made to satisfy $2^{\aleph_0} = \aleph_\alpha$ except for α excluded by Konig's theorem (that is, $\alpha = \Sigma\alpha_n$, $\alpha_n < \alpha$), the negation of the axiom of choice, and so forth [22].

The method is flexible, for example, instead of finite sets of conditions, suitable infinite ones (definable in R_{α_0}) can be profitably used [64]. In fact, by more or less ingenious applications, the *formal* character of most of the famous open problems of abstract set theory and the theory of sets of points has been settled: the answers are different for different (suitably chosen) $R_{\alpha_0}(C)$. At the time of writing it seems to be still open whether Souslin's hypothesis can be *refuted* by means of the usual axioms of set theory. For questions of *formal nonderivability* in set theory, the method is as decisive as the work of Abel and Galois for establishing the impossibility of ruler-compass (or other geometrical) constructions.

Quite recently [75] a model of set theory has been defined in which the countable axiom of choice is satisfied but all sets of real numbers are L-measurable.

Technical and Heuristic Remarks (for specialists). A more formal motivation of *forcing* (1.62) starting with 1.611, 1.612 is this. For easy comparison with the relevant literature, take D to be ω, and let $\bar{C}(n)$ denote the sequence $c_0, \ldots, c_{n-1}$ where $c_i = 0$ if $i \in C$, $c_i = 1$ if $i \notin C$. We think of C as *incompletely defined* objects. Therefore, as emphasized at the beginning of this section, the ordinary logical operations do not make sense for expressions involving C, and so we use different logical symbols; for the moment, purely formally: $(\ )$ for all; (E) for there is; $\supset$ for implies. All we assume is that, when applied to expressions not containing C, they reduce to $\forall, \exists, \rightarrow$. Let C, C' be variables for incompletely defined "sets." Then, to ensure 1.612 we put

$$1.641. \qquad A_i(C) \supset (Ep)(C')[\bar{C}(p) = \bar{C}'(p) \supset A_i(C')],$$

that is, only finitely many values of C should be *involved* in each A_i,

and

1.642.
$(C)(En)A(C, n) \supset (Em)(C)(En)(C')[\tilde{C}(m) = \tilde{C}'(m) \supset A(C', n)]$,

that is, *existential* statements should require *uniformity*. (1.611) will be automatically satisfied as long as each $\tilde{C}(p)$, for fixed p, is definable.

Now, 1.641, 1.642 are, formally, precisely the axioms 5.1 and 5.2 for absolutely free choice sequences [47]. Each condition P is determined by a sequence of 0, 1; $P \Vdash A$, A containing C, is: $(C)[\tilde{C}(p) = P \supset A]$ where p is the length of P. As usual, $\bar{A}$ may be replaced by $A \supset 0 = 1$. Then

1.643. Theorem 4 (iii) of [47] establishes the forcing rules 1.621–1.625.

In other words, the meaning of formulae $(C)[\tilde{C}(p) = P \supset A]$, for fixed P, is expressed in terms of expressions which do not contain the variable C.

Not only are the usual logical operations meaningless if applied to incompletely defined objects, but 1.641 (and 1.642) are *formally inconsistent* if regarded as statements of classical logic: take $(Ex)(y) \sim [C(y) \& \sim C(x)]$ for A_i (where $\sim B$ is short for $B \supset 0 = 1$). Conversely, it is easy to verify that if a formula A is provable by Heyting's rules, $P \Vdash A$.[9]

1.65. As a counter part of 1.535 (same notation). If a statement $\forall S \exists T \forall U A(S, T, U)$ is provable, for example, from the negation of the continuum hypothesis, it is provable outright. But there is an asymmetry between the argument here and in 1.535: there the truth of $\exists S \forall T \exists U A(S, T, U)$ was inferred from the truth in the ramified hierarchy, here one considers different $R_{\alpha_0}(C_0)$ for different S.

[9] The *notion* of forcing is not a special case of the intuitionistic notions considered in Section 2: it so happens that, for the languages considered, the usual formal laws of intuitionistic logic apply also to forcing. [If we consider *fixed* ranges of quantifiers, as above, an additional rule is valid, namely the ω-rule: if $P \Vdash \overline{A(0)}$, $P \Vdash \overline{A(1)}$, . . . then $P \Vdash \overline{\exists x A(x)}$ without the corresponding intuitionistic requirement that the premise be constructively established.] It is known [11] that the intuitionistic formal laws hold for certain elementary parts of the theory of open sets in topological spaces. The fact that they hold for forcing is more useful (or that the forcing relation can be "topologized") because nobody is interested in the elementary part of the theory of open sets per se.

1.66. Similarly to 1.65: the (finitist) relative consistency proof of [22] is (necessarily) more sophisticated than [5]; in [5] a model R_α could be defined once and for all, and each axiom of set theory shown to hold when relativized to this R_α. In [22] definitions of different $R_\alpha(C_n)$ have to be given and then one shows that the first n axioms can be formally proved to be satisfied in $R_\alpha(C_n)$.

1.67. *Consequences.* The work of [22] differs from the usual formal independence proofs in logic in being a *pure* undecidability result, in contrast to a proof envisaged in [35], cf. end of 1.5. Less hypothetically, consider Zermelo's set theory (even with the axiom of infinity); one can express the statement: there are sets not of type $\omega + \omega$. This is undecidable because the sets of type $\omega + \omega$ satisfy the axioms with the sets of type $\omega + \omega + \omega$. But once observed, this omission is regarded as an oversight, and the same is true for axioms asserting the existence of inaccessible ordinals. It may be of axiomatic interest to show that certain results hold also for sets of low type, but once one has formulated the cumulative type structure one has no reason to stop at $\omega + \omega$.

So, the problem is to find new axioms of set theory. The problem has always existed (particularly before current axioms were found!). [22] shows that one *really* has to do so even to decide the continuum problem, while previously there was only moral certainty; cf. [35] which is almost wholly devoted to explaining the nature and importance of this problem. If one were not accustomed to the reduction of mathematics to the set theoretic vocabulary with the single primitive $\in$, one would hardly expect to solve the continuum problem by use of evident axioms in this language. The problem speaks of all subsets of ω; why should not there be significant statements about such sets which simply cannot be formulated in this language, but do express something about $\mathfrak{P}(\omega)$?

The simple-minded conclusion is to look for new *primitive* notions; for example, in the marginal mathematical literature one has assertions about the "set" of *interesting* or the "set" of *feasible* numbers![10] The immediate prospect of success is not good, at least not for one accustomed to using only the set theoretic vocabulary; he finds it

[10] The system [18] contains (besides $\in$) the primitive predicate $M(x)$: x is a collection. But the axioms do not express *anything* specific about noncollections, and the intended interpretation of $x \in y$ is left open when $\overline{M(x) \cdot M(y)}$. Therefore one cannot expect the system to be strong. This is formally established in [50].

hard to take new primitive notions seriously unless they are defined set theoretically; and this is self-defeating. This is the reason why, for example, the use of infinite formulae does not seem promising in *this* connection. It is to be remarked that the search for new primitives is of independent interest. Consider, for comparison, Euclidean geometry; if we confine ourselves to the notions of betweenness and congruence we actually have a natural set of axioms which decides every question in this language [15]. But as a theory of (the concept of) space it is inadequate because it lacks the differential structure: introducing *length of curved paths* is introducing a new primitive.

A much more imaginative suggestion than the use of new primitive notions is in [35], namely to decide questions about sets of low type ($\omega + 2$ for the continuum problem) by means of

1.7. *Axioms of Infinity* (existence of sets of high type). Background: some relations between higher types and arithmetic.

1.711. In the fundamental paper [32] there are not only formally undecided statements of arithmetic, that is, with all quantifiers restricted to ω, but it is shown how simple axioms about higher types give a formal decision; for example, the consistency of Zermelo's set theory is expressed by a purely arithmetic statement, not derivable in the theory, but is derivable from the assertion that there is a set of all sets of type $\omega + \omega$.

1.712. For R_α countable in R_β, more sets of natural numbers occur in R_β than in R_α (though not necessarily in $R_{\alpha+1}$).

1.713. (In the notation of 1.53). In $R_{\alpha\mathfrak{Z}\mathfrak{F}}$ there are more sets of natural numbers than in $R_{\alpha_\mathfrak{Z}}$.

This illustrates the relevance of higher types in *principle* for deciding questions of low type. In specific cases, for example, the continuum problem $2^{\aleph_0} = \aleph_1$, there is a difficulty; for example, $R_{\alpha\mathfrak{Z}\mathfrak{F}} \cap \mathfrak{P}(\omega)$ strictly contains $R_{\alpha\mathfrak{Z}} \cap \mathfrak{P}(\omega)$, but also $R_{\alpha\mathfrak{Z}\mathfrak{F}}$ contains mappings of ω onto longer initial segments of the ordinals than does $R_{\alpha\mathfrak{Z}}$. Both $2^{\aleph_0}$ and $\aleph_1$ are increased, but by 1.511–1.513, remain equal, even on the assumption of the existence of the *usual* kinds of inaccessible ordinals. There is a generalization of the reflection principle (1.1) from which these assumptions follow, cf. [19], and which does not seem to be known to be consistent with the continuum hypothesis.

1.72. *Measurable Ordinals.* In various early attempts (for a

survey, cf. [42]) to formulate important properties of ω in set theoretic form, of which inaccessibility was one, an initial ordinal κ was called measurable provided

1.721. There is a κ-additive 2-valued measure on all subsets of κ such that points have measure 0, and κ has nonzero measure (for example, ω or 2).

1.722. $\aleph_1$-additivity (that is, countable additivity) implies that some $\kappa_1 (\kappa_1 \leq \kappa)$ is κ_1-measurable, and so $\kappa_1 > \omega$.

Measurable ordinals are inaccessible. But the basic character and the possible significance of the former were completely obscure until 1960:

1.731. Essentially any reasonable ordinal $> \omega$ is not measurable; (for methods, cf. [36], precise statements [37]) and

1.732. If $R_{\kappa'}$ satisfies the axioms of set theory, $R_{\kappa'}$ does *not* contain a measure on κ, for any $\aleph_1$-measurable $\kappa > \omega$.

The proof in [62] exploits beautifully the fact that the collecting operation of the ramified hierarchy is minimal, cf. 1.536.

A different method [60] leads to a remarkable sharpening of 1.732:

1.733. If there are measurable $\kappa > \omega$, $\mathfrak{P}(\omega) \cap R_\kappa$ is *countable;* in fact, all the ordinals accessible *in* R_κ, are countable.

[60] contains various other results, all to the effect that $\aleph_1$ is large! So, even the assumption of the existence of measurable cardinals may be insufficient to decide the continuum hypothesis. [*Added in proof:* In fact it is [74]: the cardinality of the continuum is formally not affected by assuming the existence of measurable ordinals $> \omega$.]

There is a genuine difficulty here. An axiom of infinity deserving this name must assert the existence of an ordinal which is as *different* from ω as ω is from 2 (or any other finite number). One will have to consider carefully what kind of evidence to expect for an axiom of this kind, since both exaggerated demands and too much "tolerance" make the subject trivial. A more formal problem is to give a syntactic characterization (that is, a class $\mathfrak{F}$ of formulae) of the *general form* of axioms of infinity. It seems not impossible that set theoretic truth can be *reduced* to axioms of infinity, in this sense: for each set theoretic formula A there is an A', A' in $\mathfrak{F}$, such that $A \Leftrightarrow A'$ holds in ordinary set theory (with the axiom of choice). This alone, however, would not help one to *decide* open questions since one would not know if A' is true.

1.74. *Remark on Model Theory and Recursion Theory.* In 1.73

model theoretic methods in the sense of [12] are used in an essential way, and [36] uses infinitely long formulae. The laws of logic valid for such "formulae" are expressed (i) syntactically, as (set theoretic) relations between the sets φ used as such formulae, and (ii) semantically, concerning the satisfaction relation: the structure $\mathfrak{a}$ satisfies φ. Both syntax and semantics need a generalization of recursion theory, in the sense of a theory of model theoretic invariants (cf. [56] but also [32]); thus recursion theory appears as a branch of model theory. Though model theory is a product of set theory, it is not included in the present report. It is a mathematical discipline in its own right, much of it no longer primarily concerned with foundations (of "all") of mathematics, but with a systematic understanding of particular branches:

1.741. In algebra [12] it allows a clear separation between general and specific aspects of a problem: why is the existence of an order in a commutative field equivalent to a set of inequalities? and why do we need infinitely many? In analysis, nonstandard Hilbert spaces (infinitesimals [59]) explain the occurrence of a "point" spectrum inside continuous spectra in the theory of operators; not unlike the use of the complex plane explains the behavior of power series on the real axis [for example, $\Sigma(-1)^n z^2$ diverges at $z = 1$ although $(1 + z^2)^{-1}$ is smooth there].

1.742. Besides its use as a *method*, it contains spectacular specific results such as the far-reaching generalization of the theory of real closed and algebraically closed fields in [55], and the beautiful characterization of subgroups of finitely presented groups in terms of recursion theoretic concepts in [40]. Another striking application, to $\mathfrak{p}$-adic fields, is made in the recent [73].

1.8. *Semantics and Set Theory* (self-application in the set theoretic foundations of logic). This is our first example where it is unnatural to restrict the objects considered to collections in the cumulative type structure. It is to be expected that the analysis of ordinary mathematical practice will *not force* one to introduce a more general notion of object, since (cf. page 100) much of this practice is developed from areas where there is a natural type distinction. Logic is a more promising area since it is supposed to be about arbitrary objects.

Note that the logical *operations* are part of set theory; their elimination in favor of set theoretic operations (boolean and pro-

jection) is a purely formal matter, like deciding that the chicken comes before the egg. The genuine problem concerns logical *notions* like validity, truth, consequence, and so on.

1.81. Truth, validity, and so on, in *set theoretic* structures. These notions [16] are required in model theory (1.74). To fix the ideas consider formulae of predicate logic (PL) with a single (binary) relation symbol **E**. Since all definitions are to be given set theoretically, formulae are certain finite sets (sequences of particular sets: the symbols). Objects of PL are distinguished from interpreted formulae of set theory by use of bold-face letters: lower-case letters denote (formal) variables of PL, capitals denote formulae of PL, and the (formal) constants $\wedge$, $\neg$, $\bigwedge$, $\bigvee$ are to be interpreted as $\cdot$, $^-$, $\forall$, $\exists$ respectively (the last two restricted to some set). A *substitution instance* of **A** is obtained by replacing the formal constants as intended, lower-case bold-face letters by ordinary ones, and $\mathbf{E}(\mathbf{x}, \mathbf{y})$ by some well-formed set theoretic formula $E(x, y)$.

1.811. For each fixed **A**, Sat $(x, a, e, \mathbf{A})$ [read: $x \in a$ and x satisfies **A** in the universe a with $e (e \subset a^2)$ replacing **E**] is the substitution instance $A_a(e)$ of **A** when **x** is replaced by x, all quantifiers restricted to a, and $\mathbf{E}(\mathbf{x}, \mathbf{y})$ by $\langle x, y\rangle \in e$. If **A** is closed (no free variable), this is independent of x, also written: $T(a, e, \mathbf{A})$ [read: **A** is true in the structure $\langle a, e\rangle$]. Evidently, for example

$$(\forall x a e)[\text{Sat}\ (x, a, e, \neg\mathbf{A}) \Leftrightarrow \overline{\text{Sat}\ (x, a, e, \mathbf{A})}],$$

or, for closed $\bigvee\mathbf{x}\mathbf{A}$

$$(\forall a e)[T(a, e, \bigvee\mathbf{x}\mathbf{A}) \Leftrightarrow (\exists x \in a)\ \text{Sat}\ (x, a, e, \mathbf{A})].$$

1.812. $\bigvee(\mathbf{A})$ is defined to be $(\forall a e) T(a, e, \mathbf{A})$, for closed **A** (read: **A** is set theoretically valid.) Note

1.8121. $\bigvee(\mathbf{A}) \Rightarrow (\forall a e) T(a, e, \mathbf{A})$ and

1.8122. $T(a, e, \mathbf{A}) \Leftrightarrow A_a(e)$ for each **A**.

1.82. Truth, validity, and so on, in the *universe*. There is no set theoretical formula $\mathfrak{T}(\mathbf{A})$ with *variable* **A** such that: if, for each fixed **A**, A^* is the substitution instance of **A** with unrestricted quantifiers, and $\mathbf{E}(\mathbf{x}, \mathbf{y})$ replaced by $x \in y$, then: $\mathfrak{T}(\mathbf{A}) \Leftrightarrow A^*$ (no truth definition). The reason is that on the one hand we speak of the infinite collection of PL formulae, on the other hand we use only finite formulae in set theory. What we have is this: for any fixed **A**, there is $\mathfrak{T}_{\mathbf{A}}(\mathbf{B})$ with variable **B**, such that, for fixed **B**, if **B** is not a

subformula of **A**: $\overline{\mathfrak{T}_A(\mathbf{B})}$ if **B** is a subformula of **A**: $\mathfrak{T}_A(\mathbf{B}) \Leftrightarrow B^*$, that is:

1.821. There are restricted truth definitions for the universe.

1.83. *Axiomatic Theory of Logical Validity.* The problem to be considered is the analogue of the reduction of arithmetic to set theory (1.31): one formulates laws for the basic notions, for example, logical validity: Val (**A**)—corresponding to Peano's axioms in 1.3—and looks for set theoretic conditions which satisfy these laws, or are even determined by them.

Rules of inference. Suppose certain rules have been recognized to be logically valid. The property of being derivable by the rules is set theoretically (arithmetically) definable, say $P(\mathbf{A})$ (for variable **A**).

1.831. $\forall \mathbf{A}[P(\mathbf{A}) \Rightarrow \mathrm{Val}\,(\mathbf{A})]$ is assumed to be recognized (an *axiom*).

1.832. $\forall \mathbf{A}[\mathrm{Val}\,(\mathbf{A}) \Rightarrow P(\mathbf{A})]$ expresses *completeness.* (It is rare that this is *obvious* except for a narrow class of formulae.)

1.833. $\forall \mathbf{A}[\mathrm{Val}\,(\mathbf{A}) \Rightarrow \vee(\mathbf{A})]$ is *evident* since logical validity implies set theoretic validity. The converse is problematic since Val (**A**) demands validity in all structures, not only in set theoretic structures.

1.834. $\forall \mathbf{A}[\vee(\mathbf{A}) \Rightarrow (\forall e \subset \omega^2) T(\omega, e, \mathbf{A})]$ if ω is in the class of structures considered (of course, not without the axiom of infinity; e.g., if $\vee(\mathbf{A})$ refers to validity in the class of hereditarily finite sets!).

The basic completeness theorem for the *usual* rules depends on these facts: $\forall \mathbf{A}[(\forall e \subset \omega^2) T(\omega, e, \mathbf{A}) \Rightarrow P(\mathbf{A})]$ (as a pure "set" theoretic theorem, not using regularity or the power set axiom, but using the axiom of infinity), whence, by 1.834 and 1.833, 1.832. *The problematic converse of* 1.833 *is not needed.* By the evident axioms 1.831, 1.833 the purely set theoretic: $\forall \mathbf{A}[P(\mathbf{A}) \Rightarrow \vee(\mathbf{A})]$ follows, and may be checked set theoretically. Thus

1.835. *The axioms* 1.831 and 1.833 *determine* Val (**A**) *completely to be* $\vee(\mathbf{A})$.

It is a splendid piece of luck that, for the important class of PL formulae, validity in arbitrary structures is reduced to validity in ω by these evident axioms: a priori it might have depended on the existence of measurable ordinals $> \omega$! Note that the completeness proof is conclusive only if stated for the axiomatically introduced Val, and not the defined $\vee$.

1.84. *Axiomatic Theory of Logical Validity (continued);* truth in the *universe.* Another schema of laws for the notion of validity: for each fixed **A** (cf. 1.82)

1.841. Val (**A**) $\Rightarrow A^*$, equivalently, by 1.835, $\vee(\mathbf{A}) \Rightarrow A^*$, provided only A^* is meaningful. The verification is trickier.

1.8411. The *obvious* proof by specialization of $\vee(\mathbf{A})$ runs: take the universe V for a, the membership relation $\in$ for e in $\forall a e T(a, e, \mathbf{A})$! But certainly neither is a set in the sense of the cumulative hierarchy. We should need: $\mathrm{V} \in \mathrm{V}$, that is, *self-application.*

1.8412. For *finitely* axiomatized set theories **S** (in which their own consistency, that is, $\overline{\mathrm{Val}\,(\neg\mathbf{S})}$, cannot be proved), there is an **A** for which 1.841 *cannot* be proved; take $\neg\mathbf{S}$ for **A**, since S^* is the (interpreted) conjunction of the axioms.

1.8413. In Zermelo's set theory (without regularity and without the axiom of infinity) we have, for each fixed **A**, $P(\mathbf{A}) \Rightarrow A^*$ (and hence, with the axiom of infinity, 1.841 by 1.834). The proof uses so-called *cut free rules;* cf. Section 3 and 1.821.

There is a second proof using the reflection principle: if $\overline{A^*}$, there is a *set* a such that $\overline{\exists e T(a, e, \neg\mathbf{A})}$, and thus $\overline{\vee(\mathbf{A})}$. This is the natural proof if A^* is intended as an assertion about sets in the sense of the cumulative type structure. But if A^* is to be about arbitrary objects, the argument is inconclusive because then the reflection principle is not evident. The matter is somewhat scholastic because the meaning of $\in$ itself is then not evident (cf. footnote 10).

1.842. The result 1.8413 establishes the *adequacy* of Zermelo's axioms for treating (certain properties such as 1.841 of) the intuitive notion of logical validity for the class *PL* of formula.

1.9. *Remark on Categories Which Present Two Distinct Problems.* First, a primarily formal question: can a more elegant exposition be given by use of primitives other than set and membership, even if the new ones are definable? This seems not unlikely since the fundamental notion of morphism is construed as a trinity consisting of function, domain, and "possible" range; introducing the last is like distinguishing between the empty set of apples a and the empty set of pears p, and then reintroducing extensionality by regarding them as $\langle \emptyset, a \rangle$, $\langle \emptyset, p \rangle$ respectively. Such a formal discovery might even be relevant to 1.6 if it turns out that axioms which are evident for the new primitives are not derivable from the

usual set theoretic axioms. This might happen even if the new primitives have been defined set theoretically, provided people learn to use them without constantly going back to the definition. Second, the broad question of self-application, when, loosely, there is a category of all categories satisfying a condition P, which itself satisfies P: This is nothing new in logic (1.8411), nor in mathematics. Thus the most natural way of thinking of an ordinal is as an abstract well-ordered structure (order type), and the natural ordering of these abstract structures is again a well-founded structure O. That this much makes sense is clear, less so what principles valid for the cumulation type structure are false for abstract structures. Candidates are the axiom of choice for abstract structures or the isomorphism between each well-ordered structure of abstract ordinal α and the predecessors of α in O, used in the Burali-Forte paradox.

It is amusing to observe a connection between 1.8413 and the reduction of the theory of categories to set theory adopted by its leading exponents. They use the axiom of universes, that is, the reflection principle stated for the (infinite) conjunction of all axioms of Zermelo-Frankel's set theory. Now, even without knowledge of the details, it is morally certain that the additional axiom is not needed. In any particular proof (involving categories) only a finite number A_F of axioms of set theory are used. One has almost certainly overlooked the fact that for each such case the reflection principle is provable in set theory (by use of axioms other than A_F) provided regularity is assumed. So the ideas of the reduction can be applied to the set in which these axioms hold instead of applying them to the universe of all sets.

This situation illustrates 1.8; so far mathematical practice does not *force* one to consider notions more abstract than those of the cumulative type structure. This, by itself, does not support the conclusion that therefore such notions are irrelevant to mathematics. It is common experience (cf. Section 3) that the first uses of potentially powerful principles make the exposition clearer, but can be eliminated; for example, for a long time arithmetic remained constructive, although the principle of induction permits nonconstructive uses; and even to this day analysis is exaggeratedly predicative, that is, uses surprisingly elementary instances of the principle of the least upper bound.[11] There may indeed be a reason

[11] The conclusion may be sound if one wants to use logical theory so to speak: commercially, as a means of sanctifying (defects of) current practice.

why self-application should be excluded (at least) from (realist) mathematics; but if so, this reason is not understood.

2. INTUITIONISTIC MATHEMATICS

The reader is certainly familiar with some examples of constructive and nonconstructive proofs. Almost all school mathematics is constructive; specifically, it can be codified in such free variable systems as primitive recursive arithmetic. The nonconstructive proofs that will come to his mind are probably so-called pure existence theorems where it is asserted that (for each x) there is a y with some property P, and one does not know how to find such a y. In many cases one can actually establish that there is *no algorithm* of a general kind (for example, definable by means of recursion equations) for obtaining y. Thus, in the case of the word problem for groups, one considers: for each word x there is a y, $y = 0$ and x is the identity or $y = 1$ and x is not the identity for the group considered. In analysis the situation is even simpler if one can show that y is *not* a *continuous* function of the real variable x, since then y cannot be computed approximately from approximations to x. This kind of nonconstructivity cannot even be *expressed* in free variable formalisms since they do not contain existential symbols, but consist of arithmetic identities. Though in principle in the proof of an identity nonconstructive steps, in fact nonconstructive assertions, may occur, this seems to be avoidable in current practice. It so happens that when originally nonconstructive proofs were made constructive [for example, Littlewood's theorem on changes of sign of $\pi(x) - li(x)$ or Artin's solution of Hilbert's seventeenth problem], the resulting proofs could be formulated in primitive recursive arithmetic. *These examples are too crude to serve as typical illustrations for understanding intuitionistic mathematics.* For, the negative cases are so evidently nonconstructive that $y(x)$ cannot even be defined in the required manner, let alone be proved constructively to satisfy P for each x. And the positive cases are so evidently constructive that they can be obtained by quite elementary combinatorial considerations, so-called *finitist* proofs (to be considered in Section 3). A distinction between *finitist* and *intuitionist*, first made in [34], is this: in the former, constructions are applied only to (concrete, that is, spatio-temporal) configurations, in the latter also to abstract objects (such as functions and functionals, and, particularly, "logical" operations

on so-called undecided properties and on proofs). Finitist proofs cannot be relied upon as examples of intuitionistic mathematics, because one might have quite a clear idea of constructions applied to concrete objects, and none of the more general constructions: the former idea is *independent* of the latter. Just as the concept of *finite collection* is independent of the general concept of collection.

The introduction below is intended to lead straight to a general theory of constructions. If the reader wants some typical *examples* of nonfinitist constructive proofs, he should look at 2.6 (intuitionistic mathematics of inductive definitions and ordinals) first. They are incidentally mathematically attractive; cf. 2.6411. For a discussion of nonfinitist constructive *definitions*, see 2.72–2.724; the point is important because the laws of intuitionistic logic are not particularly evident for finitist constructions. Background actually assumed: an occasional glance at some set of formal rules of intuitionistic predicate logic and of primitive recursive arithmetic. The early chapters of [8] contain this and more.

The mathematical theory begins in 2.2.

Intuitionism is a very narrow version of the idealist conception of mathematics, not unlike solipsism within general idealist philosophy. It seems to be the only version that has been developed.[12] As is to be expected from the solipsist tradition, the criticisms of rival positions are extravagant and unconvincing—and possibly merely intended to cover up the real difficulties in formulating a solipsist position coherently. But it is to be remarked that this position is quite plausible, at least as a first approximation, when one is interested in questions of *evidence* (or in the different question of intelligibility). The solipsist position stresses the *particular* evidence of those ideas which are themselves about other ideas or, more particularly, about mental acts, and not about external objects. The question is to what extent this can be developed. (Cf. Section 5 concerning the compatibility of solipsist and realist positions; of course, *not* in the strict sense of page 98.) There are still diffi-

[12] There are, of course, material difficulties in stepping from the idea of a thing to the thing itself. But are there significant *formal* (mathematical) differences between them? A possible case arises in connection with the cumulative type structure which, on the theory presented in Section 1, is not a well-defined thing (collection), and therefore unrestricted quantification over all sets was excluded in 1.1 (a hidden parameter α_0 being required). If the idea of the whole hierarchy is regarded as well defined, unrestricted quantification is well defined. But *here* the reflection principle nullifies the effect!

culties in a really coherent formulation, but there is no doubt of the intrinsic interest of the work done in this connection. This is presented below.

2.1. *General Plan.* Suggested by the experience in foundational research described in Section 1, the principal stages in the general plan are these:

Formulation and (partial) *axiomatization* of various branches of informal constructive mathematics; (i) abstract constructions and proofs (2.21 and 2.22 below); (ii) (constructive) functions of finite and higher types (2.4); (iii) free choice sequences of natural numbers and other objects (2.5, 2.62); (iv) ordinals introduced by so-called generalized inductive definitions (2.61); and finally (v) the logical operations themselves and (undecided) properties (called *species* to distinguish them from decidable properties) defined by them.

Relations between these branches in three distinct senses: (i) outright definition of one concept in terms of others, familiar, for example, from the conversion of implicit definitions into explicit ones; (ii) contextual definitions (for example, 1.31 or 1.32); and (iii) proof theoretic reductions or models (for example, 1.52). The most important examples of (ii) concern free choice sequences and generalized inductive definitions, and of (iii) concern reductions of higher type theories to arithmetic or analysis.

Finally, *isolation of primitive concepts*, in terms of which the others can be defined, and laws (axioms) for these primitives. Current candidates are *construction* (function) and the *application operation* with *proof* as a suppressed parameter.[13] It will be seen in 2.2 that these primitives are sufficient to *express* requirements, just as Peano's axioms formulated in the language of sets and relations, express what properties a set N and a relation $\subset N^2$ must have to serve as the structure consisting of the natural numbers and successor relation. But, in contrast to Section 1, we do not yet know simple axioms, valid for these primitives, whereby objects can be defined to satisfy the requirements. This is the principal problem (cf. 2.8).

A summary of proof theoretic results is given in 2.63, which exhibit remarkable interreducibility between the various branches, as developed so far. A more general way of looking at these results

[13] As *ordinal* and *order* of the cumulative type theory are suppressed in the practice of set theory. The occurrence of such hidden parameters seems essential in work that aims to give an *analysis* of informal mathematics.

in terms of the distinction between abstract and mechanical constructions is given in the discussion of Church's thesis in 2.7.

The following general observations concern *common misunderstandings* of the intuitionistic position.

2.11. *Logical Operations.* In contrast to Section 1, their meaning (particularly of quantification and implication) is genuinely problematic for the present (solipsist) approach. Thus, we make assertions about *all* numbers; something here and now is supposed to constitute evidence for this. Whatever else is in doubt, mathematical practice makes it clear that this something is (a comprehended) *proof.* So *proof* must enter into the basic explanation.

2.12. *Proofs.* For the present position, a so-called *formal* proof (sequence of symbols constructed according to a mechanical pattern) is a *representation* or *description* of a proof, and not the proof itself; just as a symbolic *description* of a set, for example, $\{\emptyset, \omega\}$, is not generally the set described.

2.121. In particular, formal rules do not express the intended meaning of correct inference, but are used only because they have been *recognized* to be valid (cf. 1.832). It should be noted that currently formal rules (of inference) are often given in the first chapter of a mathematical text, for example, Bourbaki, for the sake of *precision;* but this is demanded by a particular philosophical concept of precision (cf. Section 3) and not by mathematical practice; if it were, the later chapters of the text would have to refer back to the rules, and they do not.

2.122. To formulate (partial) laws about proofs we shall use formal systems, as in Section 1. It may so happen that the proofs represented by the formal derivations in the systems themselves also satisfy the laws. [Just as, for some set theoretic laws, the sets explicitly definable (on the basis of these laws) constitute a model.] But this is not necessarily so, for example, 2.61.

2.123. It so happens that in set theory the equality relation between sets (and the axiom of extensionality which is relevant) is quite basic. In intuitionistic *mathematics* this is less so, and often so-called intensional equality is more important. In particular, for *proofs,* equality is seldom considered and therefore one often speaks of provability.[14]

[14] But there may be an interesting theory: mathematicians interested in priority questions often discuss, in apparently objective terms, whether or not two proofs are the same.

2.13. *Proofs of What?* If the logical operations, in terms of which the usual assertions are built up, are not primitive but explained, then the *basic* proofs must be proofs of special assertions in which the (problematic) logical operations are not involved. One such special kind is familiar from quantifier-free mathematics, for example, quantifier-free arithmetic, which consists of assertions $A(n)$ for *decidable* numerical properties $A(n)$. In intuitionistic mathematics this is generalized. Instead of numbers, we have arbitrary *mathematical objects*, that is, (ideas of) concrete objects and constructions; instead of numerical properties, we have *notions*, that is, understood, decidable properties of mathematical objects. A notion is denoted by a symbol ν, and $\nu(a) = 0$ if the object a has ν, $\nu(a) = 1$ otherwise; for example, the property of being a *finite* sequence of 1 is a notion. Since notions are decidable, the usual truth functional operations may be applied to equations $\nu(a) = 0$, $\nu(a) = 1$.

2.131. *Type Structure.* [(0) is a finite type, and if σ, τ are finite types, so is σ^τ.] For each notion ν we introduce the *finite type structure* on ν as follows: $\nu_{(0)}(a) = \nu(a)$. Suppose ν_σ, ν_τ are already introduced, and $\nu_\sigma(a^\sigma) = 0$, $\nu_\tau(a^\tau) = 0$ for some a^σ, a^τ. Let ξ stand for σ^τ. For any given object x, associate $x^{(\xi)}$ by the definitions:

If
$$\nu_\tau(a) \neq 0, \qquad x^{(\xi)}(a) = a^\sigma$$
$$\nu_\tau(a) = 0 \cdot \nu_\sigma[x(a)] = 0 \Rightarrow x^{(\xi)}(a) = x(a)$$
$$\nu_\tau(a) = 0 \cdot \nu_\sigma[x(a)] \neq 0 \Rightarrow x^{(\xi)}(a) = a^\sigma.$$

Then $x^{(\xi)}$ is of type σ^τ by *definition*, and we have a well-defined notion ν_ξ of objects which are of type ξ by definition. For, if x is any clearly conceived object, it is either conceived to be introduced in this fashion or not! It is possible, of course, that $\nu_\xi(x) \neq 0$, but that we have a proof asserting: $\nu_\tau(a) = 0 \Rightarrow \nu_\sigma[x(a)] = 0$.

2.132. *Extensions and Contractions of Functions.* Given a notion ν and an object $a_0 : \nu(a_0) = 0$, and an operation b *defined* on ν. Then b can be extended to an everywhere defined operation b^* with values in ν by the stipulation: $b^*(a) = b(a)$ if $\nu(a) = 0$, $b^*(a) = a_0$ if $\nu(a) = 1$, where "if" is applied to a decidable condition. Conversely, if b is everywhere defined, we contract it to $b^0 : b^0(a) = b(a)$ if $\nu[b(a)] = 0$, $b^0(a) = a_0$ otherwise; then b^0 is ν-valued *by definition*. Naturally, the explanations above (of being of type ν_τ by definition) will not be reproduced *within* the formal development, but embodied

in the formation rules and axioms to be given. In this case, the use of truth functions applied to equations between terms of so-called higher type [34].

2.14. *A Fundamental Principle Implicit in 2.13. The meaning of* ρ is a proof of $v(a) = 0$ for all a, *is understood for notions* v; in other words, we recognize a proof of an assertion of this form when we see one. Writing $\pi[\rho; x \cdot v(x)]$ ("$x\cdot$" to denote *binding* of "x") for this relation, we have

2.141. If v is a notion, so is $\pi[\rho; x \cdot v(x)]$ where there may be other variables in $v(x)$, and x may be a sequence, not only a single variable.

This principle is embodied in the usual formal systems where, for any particular (representation of a proof by a) sequence of symbols, it can be decided whether it proves (the assertion expressed by) any given formula. In addition, formal systems require the decision (i) to be *mechanical*, and (ii) for arbitrary formulae (not only universal ones). Below we shall derive (ii) by means of 2.141 from our interpretation of logical constants; we do *not* assume (i), although most *known* (partial) laws about proofs can be shown to be compatible with (i).[15]

2.15. *The Meaning of the Primitive Concepts.* It is not to be expected that we have a clear idea of the *extension* of the concept of mathematical object; 2.132 shows that doubts about the extension need no affect particular conclusions.[16] But the application operation requires a careful interpretation.

2.151. *Total Functions.* By 2.13 we want our functions to be total, that is, all terms to denote some object, since otherwise equality ($t = s$) would not be decided. Then one would have the problem of discovering the logical laws for partially defined predicates, which (by 2.13) we wish to avoid. Now, the ordinary meaning of application $a(b)$ has to be extended; for example, if a is conceived as a function of two arguments and b as a single object, and not a

[15] This remarkable discovery, including various completeness results, constitutes much of the support of the mechanistic (formalist) position; cf. Section 3. And this has since been used in the sense of footnote 11.

[16] The difficulty is rather to find *strong* axioms for the primitive ideas. The situation is analogous to the state of set theory before the meaning of set in the sense of the cumulative type structure was clearly recognized. Such fundamental principles as the power set construction or the axiom of choice were then genuinely problematic; doubts about the extension of the concept of set were less important, cf. footnote 12.

pair, no meaning is assigned to $a(b)$. This can be met by construing every function as having only one argument [2]; but, generally, ambiguities are avoided by the following convention: if for the objects a, b as *given* or *conceived*, no sense is assigned to $a(b)$, then $a(b)$ is put $= a$, say. (Cf. in type theory: if no sense is assigned to $a \in b$, we regard $a \in b$ as false.) Obviously, for any proposed axiomatic scheme one has to verify its validity for this convention. This is illustrated by considering

2.152. *The* λ*-Calculus.* The naive proposal (parallel to the principle: every property defines a collection) is this. For every term $t[x]$, built up by means of the application operation from constants and containing the variable z, there is a function $\lambda z t[z]$ for which we have a proof $a_t : x \cdot (\lambda z t[z])(x) = t[x]$. This is excluded by the *paradoxes*. There is a notion η, $\eta(x) = 0$ if $x \neq 0$, $\eta(x) = 1$ if $x = 0$; it is a notion since any clearly conceived object either is conceived as 0 or not. Consider the term $\eta(x(x))$; though, by the convention of 2.151 it is well defined for each x, is there a clearly conceived object c (c: for Church) with $c(x) = \eta(x(x))$? No, since $c(c)$ and $\eta(c(c))$ are different. In short, the existential assumptions implicit in unrestricted λ-term formation and conversion are not correct. The "rule" c: for each x, take the value $\eta(x(x))$ overlooks the tacit convention that, for $x = c$, the value is also $c(c)$.

2.153. The two abstract theories in 2.2 below should be compared with two kinds of set theoretic formalisms. In the first, there are no λ-terms, corresponding to set theories where there are no comprehension principles at all, but only class formation operations (intersection, domain, complement, and so on); similarly, here we only have some basic notions, and *combinators*, permitting the formation of sequences, substitution operations and the like. In the second, proofs and functions are separated, though it is not excluded that proofs are also functions. Here λ-terms are used for introducing functions but with a restriction on types; only quite elementary existential assertions about proofs are made, essentially this: if, for a function constant ϕ, $\phi(a) = 0$ with free variable a has been derived by formal rules, one asserts that the very thought a_ϕ described by this formal derivation satisfies $\pi[a_\phi; x \cdot \phi(x)]$.

2.2. *Two Theories.* The first is implicit in [39]; the second is adapted from [34].

2.21. Lower-case letters, a, b, . . . (x, y, . . .) are (dummy)

variables for *constructions;* and constants if they have subscripts, or 0, 1; Greek lower-case letters are predicate constants (with possibly several arguments) for *notions;* truth functions $\cdot$, $\bar{}$ (and $\Rightarrow$ as abbreviation); $=$ for (intensional) equality; the *main* functor is *application* $a(b)$; some functors for forming ordered pairs D, $D(a, b)$ being written: $\langle a, b\rangle$, and inverses D_1, D_2, $D_1(a)$ sometimes written as a^1, $D_2(a)$ as a^2; more elegantly, one can use *combinators* (in the sense of [3]).

2.211. Terms: individual constants and variables; if t and s are terms, so are $t(s)$, $\langle s, t\rangle$, s^1, s^2. Usual notion of *free* variable.

2.212. Formulae: For terms t and s, $t = s$ (equations); if ρ is a predicate symbol (with n arguments) and $t_1, \ldots t_n$ are terms, $\rho(t_1, \ldots, t_n)$ is a formula; if α and β are formulae, so are $\bar{\alpha}$ and $\alpha \cdot \beta$. Finally, if α is a formula which does not contain x as a dummy variable and the term t does not contain the variable x, $\pi(t; x \cdot \alpha[x])$ is a formula; and x is a dummy (bound) variable. Here $\alpha[x]$ denotes α itself, and $\alpha[t]$ below will denote the formula obtained by substituting t for x in α. $\pi(a; xy \cdot \alpha[x, y])$ stands for $\pi(a^1; x \cdot \pi\{a^2(x); y \cdot \alpha[x, y]\})$.

2.213. Axioms: Equality axioms,

$$\langle a, b\rangle = \langle c, d\rangle \Rightarrow (a = c) \cdot (b = d),$$

$D_1(\langle a, b\rangle) = a$, $D_2(\langle a, b\rangle) = b$, all propositional identities.

$\pi(a; x \cdot \alpha[x]) \Rightarrow \alpha[t]$ for all terms t (reflection principle) and if $\alpha \Rightarrow \beta$ has been formally derived and the variable x does not occur in α

$$\alpha \Rightarrow \pi(a_p; x \cdot \beta[x]).$$

The subscript p can be taken to be, for example, the formal derivation itself. Free variables are suppressed, a_p depending on the variables occurring in $\alpha \Rightarrow \beta$ other than x.

2.214. *Rules of Inference.* Substitution and modus ponens. (The reader should verify, for example, the construction of a constant α_τ: $\pi(a; xy \cdot \alpha[x, y]) \Rightarrow \pi(a_\tau(a); yx \cdot \alpha[x, y])$, by use of the two principal axioms.)

2.215. Below we shall mention, but not use, the *reducibility hypothesis:*

If a notion $\rho(a)$ has a constructive characteristic function, then so has $\pi[a; x \cdot \rho(x)]$ (and similarly with more variables).

More formally; let $\alpha(c, x)$ denote the notion of application:

$c(x) = 0$. We introduce a combinator C (whose arguments and values are constructions) with the axiom:

$$\pi[a; x \cdot \alpha(c, x)] \leftrightarrow (Cc)(a) = 0.$$

For important consequences, see 2.6142. If in some sense, constructive functions can be $\neq 0$ only for some a priori limited kind of argument, the reducibility hypothesis means this: there is some a priori limit on the kind of construction which can be used as a proof about any given constructive function.

This completes the description of our first system.

2.22. Lower-case letters with type superscripts a^τ, b^σ, . . . are variables for functions of type $\tau, \sigma, \ldots$; and constants if they have subscripts; German lower-case letters for proofs; a single relation symbol = between terms; and $\pi(\mathfrak{a}; a^\tau b^\sigma \ldots \cdot \alpha[a^\tau, b^\sigma, \ldots])$ for α which are propositional combinations of equations. *Application* $t^\sigma(s^\tau)$ if σ is of the form ρ^τ.—λ-symbols.

2.221. Terms are built up in the usual way subject to type restriction.

2.222. Formulae: Propositional combinations of equations, or $\pi(\mathfrak{a}; a^\tau b^\sigma \ldots \cdot \alpha[a^\tau, b^\sigma, \ldots])$.

2.223. Axioms: Equality axioms, propositional identities, and, for each λ-term, the schema of λ-conversion. (λ-terms can be replaced by suitable combinators, just as in classical set theory the comprehension schema can be replaced by use of a finite number of class operators.)

2.224. Rule of inference: Substitution.

Also, if $\alpha[a^\tau, b^\sigma, \ldots]$ has been formally proved for free variables $a^\tau, b^\sigma, \ldots$

$$\pi(\mathfrak{a}_p; a^\tau b^\sigma \ldots \cdot \alpha[a^\tau, b^\sigma, \ldots]).$$

2.23. The axioms given above are the *basic* axioms (for the foundation of logic); for example, if type 0 variables in the second theory are intended to be number variables, the principle of definition by recursion is added by introducing a successor constant $'$, and, for given constants b_1 c_1, a constant a_1 with the axioms:

$$a_1(0) = b_1 \qquad a_1(a') = (c_1(a))(a_1(a))$$

where b_1 is of type τ, a_1 of type τ^0, therefore a of type 0, and c_1 of type $(\tau^\tau)^0$; cf. [34].

2.3. *Two Meanings of Logical Operations; Validity of Formal Rules of Predicate Logic* (hence: reduction of logical primitives to the primitives of 2.2).

Formulae in the notation of predicate logic are considered, built up from predicate symbols P_i^n with n arguments.—N.B. Though the meanings are different in the two cases we use the same symbols: $\sim$ (not), & (and), $\vee$ (or), $\supset$ (implies), (x) (for all x), (Ex) (there is an x).

2.31. Throughout we show at most one free variable. Each predicate symbol is assumed to be *interpreted*, that is, it is understood what it means that a is a proof that x satisfies P_i, that is, there is given a notion $\rho_i(a, x)$. By induction one builds up the notions expressing: a is a proof of a (complicated) formula $A(x)$, denoted by $\Pi[a; A(x)]$:

$\Pi[a; P_i(x)]$ is $\rho_i(a, x)$; $\Pi(a; A \,\&\, B)$ is $\Pi(a^1; A) \cdot \Pi(a^2; B)$,

$\Pi(a; A \vee B)$ is $\overline{\overline{\Pi(a^1; A)} \cdot \overline{\Pi(a^2; B)}}$; $\Pi[a; (Ex)A(x)]$ is $\Pi[a^1; A(a^2)]$;

$\Pi[a; (x)A(x)]$ is $\pi(a^1; x \cdot \Pi[a^2(x); A(x)])$;

$\Pi[a; A \supset B]$ is $\pi[a^1; x \cdot \Pi(x; A) \Rightarrow \Pi(a^2(x); B)]$.

Evidently, for each well-formed expression $A(x)$ of predicate logic, $\Pi[a; A(x)]$ is a formula of 2.21 built up from π and ρ_i. We have

2.311. *A formula A of predicate logic is formally derivable by means of the usual rules, if and only if for some constant a_A of* 2.21, $\Pi(a_A; A)$ *is a theorem of* 2.21.

2.312. In contrast to 1.831 and 1.832, 2.311 is stated as a proof theoretic equivalence, and not as an implication. The latter case will be considered in 2.74 (completeness problem).

2.32. For the second meaning, each predicate symbol is assumed to be interpreted in the form: it is understood what it means that $\langle a, a^x\rangle$ is a proof that x satisfies P_i, here given by a notion $\rho_i(x, a^x, b^\sigma)$ (a and b are *sequences*), such that $\pi[a; b^\sigma \cdot \rho_i(x, a^x, b^\sigma)]$. If P_i is decidable (a notion), ρ_i is independent of a and b.

We drop type superscripts; the understanding is that the types used are coherent. Though a variable a for proofs is shown, for easy comparison with 2.31, note that *no constructions on proofs are assumed* except for trivial pair formation and specialization.

We write $\Gamma(\mathfrak{a}; a; A)$; each such Γ will have the form:

$\pi(\mathfrak{a}; b \cdot \alpha[a, b])$ (using $\pi(\mathfrak{a}; b \cdot \alpha[b]) \cdot \pi(\mathfrak{a}'; c \cdot \beta[c])$
$\Leftrightarrow \pi[\langle \mathfrak{a}, \mathfrak{a}' \rangle; bc \cdot \alpha(b) \cdot \beta(c)]$

$\Gamma[\mathfrak{a}; a; P_i(x)]$ is $\pi[\mathfrak{a}; b \cdot \rho_i(x, a, b)]$,
$\Gamma(\mathfrak{a}; a; A \ \& \ B)$ is $\Gamma(\mathfrak{a}^1; a^1; A) \cdot \Gamma(\mathfrak{a}^2; a^2; B)$;
$\Gamma(\mathfrak{a}; a, A \vee B)$ is $\pi(\mathfrak{a}; bc \cdot \{(a^1 = 0 \Rightarrow \alpha[a^2, b]) \cdot (a^1 \neq 0 \Rightarrow \beta[a^2, c])\})$
where
$\Gamma(\mathfrak{a}; a; B)$ is $\pi(\mathfrak{a}; c \cdot \beta[a, c])$;
$\Gamma[\mathfrak{a}; a; (Ex^\tau)A(x)]$ is $\Gamma[\mathfrak{a}; a^1; A(a^2)]$;
$\Gamma[\mathfrak{a}; a; (x^\tau)A(x)]$ is $\pi[\mathfrak{a}^1; x^\tau \cdot \pi(\mathfrak{a}^2(x); b \cdot \alpha[a(x), b])]$
Finally, $\Gamma(\mathfrak{a}; d; A \supset B)$ is $\pi[\mathfrak{a}; ab \cdot (\alpha[a, d^1(a, b)] \Rightarrow \beta[d^2(a), b])]$.

2.321. *If a formula A of predicate logic is formally derivable by means of the usual* (intuitionistic) *rules, then there is a proof constant $\mathfrak{a}_A$, and a function constant a_A such that $\Gamma(\mathfrak{a}_A; a_A; A)$ in* 2.22. *If A is derivable in arithmetic then there is a constant a_A of the system extended by primitive recursion* 2.23.

2.322. The converse of 2.321 does not hold since for A: $\{(x)[P(x) \vee \sim P(x)] \ \& \sim\sim(Ex)P(x)\} \supset (Ex)P(x)$, we have $\Gamma(\mathfrak{a}_A; a_A; A)$, but A is not formally derivable. In other words, the usual rules are not complete for the meaning given to the logical operations in 2.32. This is not surprising: if $P_i(x)$ is itself decidable, that is, $\rho_i(x, a, b)$ is independent of a and b, the interpretation of $\sim\sim(Ex)P(x)$ is just that of $(Ex)P(x)$.

2.33. *Interpreting the Interpretations of the Logical Connectives.* Evidently, the trivial axioms of 2.21, 2.22 leave wide latitude to what one is to mean by *proof* or *constructive functions;* obviously one cannot expect that the elementary practice of logic or arithmetic determine such general ideas. The examples below give some idea of alternative interpretations. (The conceptual defect of ambiguity will be turned into proof theoretic reductions.)

2.331. Alternatives for the axioms 2.22 are these: (i) constructive functions of constructive functions, etc., as in [34], (ii) additional continuity requirements (for the topology of [46]), and (iii) the modification of (ii) in [48] p. 154.

2.332. Which of these alternatives is adopted may affect the *relation* between 2.31 and 2.32. Here are some examples for arithmetic A:

2.3321. For prenex A, the two meanings are equivalent, for negations of such A, 2.32 is stronger than 2.31 provided the same functionals are considered. So here the relation is clear.

2.3322. Let A^- be the *negative* version of A [where (Ex) is replaced by $\sim(x)\sim$, $A \vee B$ by $\sim(\sim A \,\&\, \sim B)$] and $\bar{A}$ the usual classical prenex form of $\sim A$. Then for *continuous* functionals [2.331 (ii) or (iii)] of type 0^{0^*}, $\sim\sim A$, A^-, $\sim\bar{A}$ are all equivalent on 2.32, but not on 2.31. There we only have $\sim\sim A \supset A^- \supset \sim\bar{A}$ (for prenex A).

Which of these alternatives is intended will come up again in 2.5 (page 134) and above all in 4.4.

2.333. For purely formal proof theoretic results in Section 4, note

2.3331. If in 2.32, arbitrary (nonconstructive) functionals are used instead of those of 2.2, the logical connectives satisfy the laws of classical logic instead of 2.321.

2.3332. If A is a formula of finite type (not necessarily arithmetic), and A_1 the interpretation by 2.32 of A^-, then $A \Leftrightarrow A_1$ can be proved by classical logic of finite order (without comprehension axiom) by use of the axiom of choice: $\forall a^\tau \exists b^\sigma C(a, b) \Rightarrow \exists c^{\tau^*} \forall a^\tau C[a, c(a)]$ *restricted to quite elementary quantifier-free* C only. [46, p. 120].

2.34. *Axiom of Choice.* The axiom of choice $(x)(Ey)A(x, y) \supset (Ez)(x)A[x, z(x)]$ holds on both interpretations for either unrestricted quantifiers x, y, or for x, y restricted to decidable ranges. For 2.32, and decidable A,

$$(x)[(z)A(x, z) \supset (Ey)B(y)] \supset (x)(Ey)[(z)A(x, z) \supset B(y)],$$

but not generally on 2.31. Generally, for unrestricted $A(x)$: $(x)[A(x) \supset (Ey)\{A(y) \,\&\, B(y)\}]$ asserts on 2.31 the existence of a function of *two* variables $f(x, a)$:

$$\Pi[a; A(x)] \Rightarrow \Pi(f^{11}(x, a); A[f^{12}(x, a)]) \cdot \Pi(f^{21}(x, a); B[f^{22}(x, a)]).$$

An "object" in the range $A(x)$ is a pair, namely an object x and a proof a that x satisfies $A(x)$.—N.B. Even if y is known to be *unique*, one cannot generally prove by known principles that the quantifier can be brought forward (3.3221).[17]

2.35. *Definitional and Extensional Equality.* For given terms t and s (even constants), $t = s$ may be *formally* undecided; for,

[17] Let $Ev A(x, v)$ be nonrecursive, A primitive recursive; so that $(x)\{(Ev)A(x, v) \supset (E!y)(z)[A(x, y) \,\&\, \{z < y \supset \sim A(x, z)\}]\}$; if y could be brought forward every recursively enumerable set would have a constructive characteristic function.

in an interpretation one may have in mind more than just the rules of conversion explicitly set out in the axioms. *One* interpretation of 2.22 is that the operations denoted are simply those *fully described* by the conversion rules themselves (in terms of application and substitution). Though abstract thought is needed to recognize the fact that they give a full description, the equality relation (interconvertibility of terms) may be expected to be mechanical, that is, recursive. For the theory 2.22, this is verified in [67] and analyzed.[18] It should be noted that not all clearly understood rules (even for purely number theoretic functions) can be described in terms of application and substitution: consider a formal language and a meaning; if a sequence of symbols denotes a constructive proof a of an existential assertion $(Ex)A(x)$, take the object a^1 of *that* proof; otherwise take 0. Proof theory for *intuitionistic* formal systems often reduces this to a recursive rule.

In 2.21, *extensional* equality $a \equiv b$ means: $(x)[a(x) = b(x)]$. If a, b are not given as universally defined constructions, for our particular convention in 2.151 we then have $a \equiv b$ only if $a = b$.

In 2.22, this is a weaker sense; for type 0 objects, $a \equiv b$ means: $a = b$. The stronger sense is defined by induction with respect to the finite type structure. $(x)[a(x) \equiv b(x)]$ and $(x)(y)[x \equiv y \supset \{a(x) \equiv a(y) \,\&\, b(x) \equiv b(y)\}]$.

2.4. *Functions of Higher Type or Logic (Proofs)?* (Cf. 2.131.) All logical operations are to be regarded as either primitive or in the sense 2.31 (not 2.32). The present section (2.4) contains mainly *proof theoretic* interreducibility results.

2.41. To give an idea of the facts we need some definitions.

2.411. T is the quantifier free system (of primitive recursive arithmetic) of finite types of 2.321, that is, of [34], with intuitionistic propositional logic, that is, not $t = s \vee \sim t = s$ for all terms of higher type, T_D (D: for definitional), with truth functions applied to all equations, T_E (E: extensional) with the rule, from $P \supset t[x] = s[x]$

[18] So to speak: a well-defined minimal model of 2.22. It has always been evident that, except in trivial cases, no set of axioms can be expected to determine a unique interpretation of the kind above, unless supplied by some tacit convention like minimality. More surprisingly the situation has turned out to be completely parallel for general models in *classical* logic (relativity and incompleteness theorems). Except for details, the role of formal systems is much the same in Sections 1 and 2.

(only equations between terms of lowest type) infer $P \supset u(\lambda x t[x]) = u(\lambda x s[x])$.

2.412. For a set $[\tau]$ of types, $H_{[\tau]}$ is T restricted to types in $[\tau]$, with quantifiers, intuitionistic logic, induction applied to *all* formulae, and the axiom of choice applied to *quantifier free A:* if ξ, η, $\zeta(=\eta^\xi)$ are in $[\tau]$, $(a^\xi)(Eb^\eta)A(a, b) \supset (Ec^\zeta)(a_\xi)A[a, c(a)]$. H is first order arithmetic, H_1 has also variables for number theoretic functions with a finite number of arguments. These are the basic systems, more axioms will be added.

2.42. By [34], T is much stronger than primitive recursive arithmetic (ϵ_0 against ω^ω). In *the presence of logic the situation is quite different: finite types are absorbed.*

2.421. H_ω with the *unrestricted* axiom of choice (all A) is reduced to T, literally as in [34]. Conversely,

2.422. T, T_E, T_D are interpreted in $H \cdot T_D$ by [67]; for T_E, use 2.331 (ii), that is, use first H_1, define in H_1 *neighborhood functions* for the topology [46], by conditions $C_\tau{}^E(a)$, interpret (a^τ) as $(a)[C_\tau(a) \supset]$, and verify the axioms of T_E. For reference below: by modifying $C_\tau{}^E$ [2.333 (iii)], one gets an interpretation for T with *weak* extensionality of 2.35, and in addition a modulus of continuity functional for operations on sequences. Then H_1 is interpreted in H by letting a range over the recursive functions, the *basic* axiom of choice being satisfied.

2.423. For extensions to transfinite types, cf. [67]. The results have the same character for suitable types other than ω. Since the existential axioms of [67] require an ordinal α to be given the only self-contained scheme provided by [67] for introducing functions is this: a type α is introduced in the quantifier free theory if α has been proved to be a well ordering (in the obvious sense). We get to all ordinals $< \Gamma$ (3.51); by 3.531, the same is reached if one considers the (intuitionistic) ramified hierarchy of logically definable properties (species) of natural numbers, and iterates it α times if α has been proved to be a well ordering in an earlier system (cf. 2.153). Note that the reduction of 2.422 cannot be applied since, for transfinite τ, $C_\tau{}^E$ is not definable in H_1.

2.43. Two comments on the axiom of choice. First its proof theoretic strength in intuitionistic logic is much weaker than in classical logic; for instance, the addition of the law of the excluded middle to H_1 and the unrestricted axiom of choice at the lowest level gives full classical analysis, in contrast to 2.421. Second, while T_E

is satisfied by the functionals described in 2.422, H_ω with the unrestricted axiom of choice is not: apply it to $(a^2)(a^1)(En)(b^1)$ $[(m < n)(b^1m = a^1m) \supset a^2(a^1) = a^2(b^1)]$. This is quite consistent with 2.34 since the modulus of continuity n depends not only on a neighborhood function $a^{(2)}$ of a^2 but on the proof of $C_2(a^{(2)})$.

2.5. *Free Choice Sequences; objects* (to study) or *figures of speech* (to get rid of)? The *obvious* source is this: most constructive operations on functions, for example, decimal expansions, are *continuous*, using only finite sequences of values (not the rules for calculation) of their arguments. So, the lattter may be given by approximations that are thought of as freely chosen. One probably does not attach any immediate sense to logically complicated expressions with quantifiers over free choice sequences (denoted by Greek lower-case letters); cf. points of infinity in geometry, used primarily as shorthand for talking about parallel lines. If one could find a clearly conceived idea, that is, a mathematical object, "behind" this figure of speech, one would have a new *primitive* idea and possibly new axioms (about unbiased coin tossing, or impenetrable fate in Brouwer's later writings). These axioms have to isolate properties of the notion: a is a proof of $(\alpha)A(\alpha)$, $(E\alpha)A(\alpha)$ (for complicated A).[19] The implicit assumption[20] is that these properties will be coherent with the interpretation of the logical operations to which we are "committed" by 2.31.

For figures of speech the commitment is absolute: there are no free choice objects and so nothing new is involved in *proofs* and the functions that are part of the proofs. For example, suppose, for some purely arithmetic A, $\sim(n)[A(n) \vee \sim A(n)]$. Then $\sim(E\alpha)(n)\{[A(n) \,\&\, \alpha(n) = 0] \vee [\sim A(n) \,\&\, \alpha(n) = 1]\}$. This may seem to contradict the freedom left open to α (if one has unconsciously stepped into a realist or possibly formalist interpretation), but does not: for a *proof* of $(E\alpha)(n)\{\ldots\}$ would provide a proof of $(n)\{\ldots\}$and further, some particular α that decides between $A(n)$ and $\sim A(n)$.

For free choice sequences as objects, there is a fundamental problem: are they to be involved in proofs? Or are only (clearly defined) constructions *on* such objects to occur in proofs? In the

[19] It may be debated whether (i) understanding the meaning of $(\alpha)A(\alpha)$ simply means (ii) understanding this relation; all we need is that (i) implies (ii).

[20] As in the whole section: clearly conceived ideas are coherent!

former case, a is a proof of $\sim A$, means that it is a proof of $\overline{\Pi(x, A)}$, where (on the most liberal interpretation) x may involve free choice sequences even if A does not. Thus, if the variables in 2.31 denote objects which themselves involve free choice sequences, the "commitment" is pretty loose!

An analysis of this second conception, in effect, if not intended, is in the monograph [8].[21]

It verifies completely what one expects: on the basis of really evident axioms for free choice sequences one does have strong continuity properties, but $\sim A$ can be asserted essentially only if A is *false* for arbitrary functions (modulo the continuity properties); cf. 2.6. A parallel situation arises with the interpretation 2.32 if *continuity* requirements are imposed on the functionals used in interpreting the connectives, cf. 2.331 (ii).

The axioms below are clear; that is, our idea of free choice sequence may be so vague that it leaves most questions undecided, but not these. But they turn out to be reducible to arithmetic—though, (from the wrong point of view) they look inconsistent; cf. 1.643. In 2.5, for studying the main axiom (misnamed: bar induction) we seriously regard free choice sequences as figures of speech. As is to be expected when something is "analyzed away," much sharper proof theoretic consequences are obtained than in [8].

Also, and this is probably more important, *if* the main axiom is accepted, the *elimination of free choice sequences* (of constructive objects) *is a consequence*. The status of other kinds of freely chosen object, that is, not only freely chosen sequences of *constructive* objects, is probably the *principal open question* of the subject; cf. 4.31.

2.51. (Cf. 2.412). $H^{[\sigma]}_{[\tau]}$ is obtained from $H_{[\tau]}$ by extending the language and axiom schemata by adding variables α^ξ (for $\xi^0 \in [\sigma]$); they are treated as ξ^0 objects in the formation rules $(\alpha^\xi)(n)(Ea^\xi)[\alpha^\xi(n) = a^\xi]$. But if constructions are not to involve free choices, there are no constructions with α^ξ objects as *values*. What one can do is to take operations: $\xi^0 \to \eta$ (with $\eta \in [\tau]$) as new basic types, and extend the schemata of T to these as in 2.6244. Here we con-

[21] Recursive realizability [8]. In contrast to 2.3, this analysis takes all logical operations as primitive except (Ex) ($A \vee B$ is replaceable by $(Ex)[(x = 0 \supset A) \,\&\, (x \neq 0 \supset B)]$) which occurs only in the special context: $\{e\}(x)$ is defined. (x) is applied to formulae of complicated structure. Nevertheless, it is sometimes proof theoretically very useful, for example, 4.121.

sider $[\sigma] = 1$ only (sequences of natural numbers), and start with H^1, obtained from H by adding, for *all* A, the axiom of choice: $(n)(E\alpha)A(n, \alpha) \supset (E\alpha_1)(n)A_1(n, \alpha_1)$; here one uses the following convention: let $<n, m>$ be a numbering of pairs of natural numbers, and $A_1(n, \alpha_1)$ obtained from $A(n, \alpha)$ by replacing each $\alpha(t)$ by $\alpha(<n, t>)$, in short, if $A(n, \alpha)$ is $A[n, \lambda x\alpha(x)]$, $A_1(n, \alpha_1)$ expresses $A[n, \lambda x\alpha_1(<n, x>)]$. ($H^{[\sigma]}_{[\tau]}$ is analogous, but 2.412 is restricted to A which do not contain free Greek lower-case variables.)

2.511. *Continuity.* Let $C(\gamma)[C_0(\gamma)]$ express that γ is a (freely chosen) neighborhood function into the space of sequences α of natural numbers [respectively, into space of natural numbers], that is, if $\bar{\alpha}m$ denotes $\langle\alpha(0), \ldots, \alpha(m-1)\rangle$, $\bar{\alpha}0 = \langle\ \rangle$, German letters denote finite sequences of natural numbers, then

$$(n)(\alpha)(Em)[\gamma(\langle n, \bar{\alpha}m\rangle) \neq 0] \,\&\, (n)(\mathfrak{m})(\mathfrak{p})\{[\gamma(\langle n, \mathfrak{p}\rangle) \neq 0 \,\&\, \mathfrak{p} \subset \mathfrak{m}]$$

$$\supset \gamma(\langle n, \mathfrak{p}\rangle) = \gamma(\langle n, \mathfrak{m}\rangle)\}$$

[with the first variable n missing in $C_0(\gamma)$]. Let $\Gamma(n, \alpha)$ be the functional defined by γ. For all A

$$(\alpha)(E\beta)A(\alpha, \beta) \supset (E\gamma)[C(\gamma) \,\&\, (\alpha)A_2(\alpha, \gamma)]$$

where $A_2(\alpha, \gamma)$ expresses $A[\alpha, \lambda n\Gamma(n, \alpha)]$.

2.512. H^1 with 2.511 is reduced to H by 2.321 and 2.422 with *modified* $C_\gamma{}^E$. One has extensionality $(x)[\alpha(x) = \beta(x)] \supset [A(\alpha) \supset A(\beta)]$ since by the formation rules α, β enter only through their values.

2.52. To state more significant facts, we consider $H_1{}^1$ and the following *basic axiom:*

2.521. $S(a)$ [read: a is a constructive spread] means that a defines a set of permitted initial segments, with $\mathfrak{a}$ ranging over finite sequences of numbers and * denoting concatenation: $(\mathfrak{a})(Ex)[a(\mathfrak{a}) = 0 \supset a(\mathfrak{a}{}^*x) = 0]$, $(\mathfrak{a})(\mathfrak{a}' \subset \mathfrak{a})[a(\mathfrak{a}) = 0 \supset a(\mathfrak{a}') = 0]$, and $\alpha \in a$: $(x)[a(\bar{\alpha}x) = 0]$. Then, for all $A(\alpha)$ containing possibly bound, but no free Greek lower-case letters,

$$A(\alpha) \supset (Ea)[S(a) \,\&\, \alpha \in a \,\&\, (\beta \in a)A(\beta)]$$

(the converse being logically valid). For, to be *given* α is to be given a spread. It follows

2.522. $(a)\{[S(a) \,\&\, (\alpha \in a)A(\alpha)] \supset [(\alpha \in a)B(\alpha)]\} \supset [A(\alpha) \supset$

$B(\alpha)]$ (the converse is trivial), and $(E\alpha)A(\alpha) \supset (Ea)A(a)$ if there are no free Greek letters. Then (for *negative* as in 2.3322)

2.523. *First Elimination Results. If A is a negative formula, containing bound, but no free Greek variables, there is A' not containing any such variables, such that $A \leftrightarrow A'$, in $H_1{}^1$ with* 2.521

(2.511 is of course not needed). Proof by induction on number of (universal) Greek quantifiers. Hence

2.524. For quantifier free A, $(\alpha) \sim(x)\sim A(\bar{\alpha}x) \leftrightarrow (a)[S(a) \supset \sim(b \in a) \sim A(\bar{b}x)]$.

2.525. Trivially, the extension of $H^1 + 2.511$ by 2.521 is conservative by simply treating a (constructive functions) as α.

2.53. The last *elementary axiom* expresses that the *whole* extension of a *constructive* function (rule) depends only on a finite number of values of an α, in contrast to 2.511. For $A(\alpha, a)$ not containing free Greek letters other than α:

2.531. $(\alpha)(Ea)A(\alpha, a) \supset (Eb)(Ec)[C_0(c)\ \&\ (\alpha)A_3(\alpha, b, c)]$ where $A_3(\alpha, b, c)$ expresses that the function $a = \lambda nb(\langle n, \Gamma(\alpha)\rangle)$ satisfies $A(\alpha, a)$. In other words $(\alpha)(E\beta)$ requires a continuous mapping from N^N to N^N where domain and range have the product topology, $(\alpha)(Ea)$ one where the domain has the product topology and the range the discrete topology.

The trivial method of 2.525 cannot be used to reduce $H_1{}^1 + 2.511 + 2.521 + 2.531$ to arithmetic because 2.531 and 2.511 conflict. One uses again the method of 2.321.

To extend 2.523 to *all* (closed) formulae A, one has to analyze more closely statements of the form $(\alpha)(E\beta)A(\alpha, \beta)$. By 2.511 this is thrown back on $(\alpha)(Ex)A(\alpha, x)$ [in the definition of $C(\gamma)$] actually even for quantifier-free A. By 2.522, if $(\alpha)(E\beta)A(\alpha, \beta)$ contains no free Greek letters, γ may be taken to be constructive.

The upshot of the analysis is this. *On the assumption that an α is given outright by a spread,* that is, 2.521, *the notion: a* is a proof of *A, is fully determined for all A, if it is known for $(\alpha)(Ex)A(\alpha, x)$.*[22] This is in effect the subject of

2.6. *Generalized Inductive Definitions; Bar Theorem.* A formal *reduction* of: $(\alpha)(Ex)$[or: $(\alpha^r)(Ex)$], needs of course a machinery

[22] More precisely, to each formula A not containing free variables for free choice sequences, there is a formula A', equivalent to A on the basis of our axioms and built up from the predicates $C_0(b)$ of 2.511 without the use of *any* variables for free choice sequences.

which does not explicitly involve free choice sequences, and so, at first, may seem to have nothing to do with them. This requires additions[23] *either* of (axioms from which) new constants and their laws (can be derived) *or* new (basic or transfinite) types. This is of independent interest. Since in 2.2 the notion of proof is primitive, the old-fashioned idea of inductive definitions suggests itself.

2.61. *Background.* Let $A(P, x)$ with the predicate symbol P (monadic; or else n-ary and x a variable over n-tuples of constructive objects) be well formed, that is, its meaning defined once P is *interpreted.* Inductive definitions of (the interpreted species) P_A take the form

2.611. $P_A(x)$ if and only if $P(x)$ can be "proved" from:

$$(x)[A(P, x) \supset P(x)].$$

Inverted commas are used, because the notion of proof (2.2) applies to interpreted statements, while P is a formal letter; so to speak: nothing should be assumed about P. *If* a clear sense is given to 2.611, then P_A *is* interpreted and, for any (interpreted) Q, we have the *principle:*

2.612. $(x)[A(Q, x) \supset Q(x)] \supset (x)[P_A(x) \supset Q(x)]$.

A quite separate question is whether 2.611 is useful, that is, that something interesting can be said about P_A, in particular

2.613. $$(x)[A(P_A, x) \supset P_A(x)].$$

Trivially, for numerical x, 2.613 does not hold if $A(P, x)$ is

$$(Ey)[\sim P(y) \,\&\, x = y + 1].$$

2.614. What $A(P, x)$ to use? This depends of course on the precise sense in which 2.611 is understood. The situation is *exactly* parallel to the classical case.

2.6141. Reduction to definition by induction. If $\perp$ is (the) empty (species), put $P_0^A(x) \leftrightarrow A(\perp, x)$, and, for $n \neq 0$, $P_n^A(x) \leftrightarrow (Em < n)A(P_m^A, x)$. For positive A, that is, conjunctions and disjunctions of unnegated prime formulae, and $n < \omega$, each $P_n^A(x)$

[23] The need for *additions* follows from the results below since the formal laws for $(\alpha)(Ex)$ to be derived, allow one to prove the consistency of the systems in 2.2.

is a primitive recursive predicate of x, and $P_A(x) = (En < \omega)P_n^A(x)$ satisfies 2.612 and 2.613.[24]

In this simple case there are *formal* rules such that $P_A(x)$ if $\mathbf{P}(\mathbf{S}^z\mathbf{0})$ is formally derivable from $(\mathbf{x})[A(\mathbf{P}, \mathbf{x}) \supset \mathbf{P}(\mathbf{x})]$. For quantified A we do not have closure at ω, nor formal rules; that is, the sequence P_n^A is continued into the transfinite, and thus the analysis thrown back on a theory of *ordinals*. These in turn are introduced by an inductive definition: (i) In the restricted case of (primitive) *recursive* ordinal notations, the binary relation $<_O$ is inductively defined by $O(<_O; x, y)$:

$$(x = 1 \,\&\, y = 2) \vee (1 <_O x \,\&\, y = 2^x) \vee (Ez)(x <_O z \,\&\, z <_O y)$$
$$\vee (Ez)(Em)\{y = 3.5^z \,\&\, x = \{z\}(m) \,\&\, (n)[\{z\}(n) <_O \{z\}(n + 1)]\}.$$

(ii) For the general case, cf. 2.8. Thus this *reduction* of inductive definitions is nothing else but the reduction of a *class* of such P_A to a *particular* one, namely, P_O, where $P_O(x) : 1 \leq_O x$ (x is an ordinal notation). For example, for all purely universal positive A, P_A can be explicitly defined from P_O. It does not seem to be known if this holds for all monotonic A.

2.61411. Much of the literature on recursive ordinals is done nonconstructively. Closer inspection shows that *this* theory is much more elegant if done intuitionistically. For instance, for either a finitist or a nonconstructive recursion theoretic approach, it comes as a surprise that, for example, (normal ordinal) functions defined by induction on the highly nonrecursive set O turn out to be primitive recursive. (This is shown by more or less ad hoc applications of the recursion theorem.) In contrast, harmless looking functions, such as addition, on some specific segment of O, need not be recursive at all. This is exactly what one expects on Church's thesis (2.7) if one considers the constructive meaning of the logical operators (constructions on abstract objects).

2.6142. Reduction to the abstract theory of constructions requires, in the notation of 2.2, a notion $\rho_A(a, x)[a$ is a proof of $P_A(x)]$ for which 2.613 and 2.612 can be verified for the Q considered. The obvious analogue to Dedekind's approach in set

[24] In the classical case, a weaker condition on A is sufficient, namely *monotonicity;* that is, writing $P \subset P'$ for $(y)[P(y) \to P'(y)]$, $(x)\{[A(P, x) \wedge P \subset P'] \supset A(P', x)\}$. This is no longer true here; let A be positive, P_A nonrecursive, and let $B(P, x)$ be $\sim\sim P(x) \cdot \vee A(P, x)$. Then $P_B(x) \leftrightarrow P_{\omega+1}^B(x)$ can be proved, but not $\leftrightarrow P_\omega{}^B(x)$.

theory is this: a proves that for all notions P with *constructive characteristic* function p [of two variables! b is a proof of $P(x)$], $(P)(y)[A(P, y) \supset P(y)] \cdot \supset P(x)$. Just as in set theory, this notion can be *expressed* in 2.21, but need not have a constructive characteristic function; thus if Q may contain P_A, 2.612 does not follow. It does follow from the reducibility hypothesis of 2.2; this also implies essentially the comprehension axiom if it is assumed that the notion ν of being a natural number has a constructive characteristic function. It does not seem to be known whether obvious weaker conditions are sufficient to derive 2.612, and 2.613 (for proper A); perhaps one only needs that the binary relation: a is a proof that x is a natural number, have a constructive characteristic function (and not a completed set ν).[25]

2.615. Leaving for the moment the question of a more thorough constructive analysis *of* inductively defined P_A, we shall use them as a *means* of analysis; at least for positive A.[26] For it is intuitively clear that any satisfactory theory of constructivity will include them.

2.6151. Note that the principle of *proof by transfinite induction* on O is a particular case of 2.612; 2.6152: not, however, that every Q has a first element; 2.6153: $(\alpha)(Ex) \sim [\alpha(x + 1) <_O \alpha(x)]$ (trivially, that is, using only $(\alpha)(E\beta)(x)[\beta(x) = \alpha(x + 1)]$).

2.62. *The Quantifier* $(\alpha)(Ex)$. Consider a decidable property B of sequences of natural numbers and, as above, write $\bar{\alpha}x$ for the sequence $\alpha(0), \ldots, \alpha(x - 1)$ ($\bar{\alpha}0 = <\ >$). Brouwer [20] proposed that a proof a of $(\alpha)(Ex)B(\bar{\alpha}x)$ has the form:[27] either for some fixed x_0, a proves $(\alpha)B(\bar{\alpha}x_0)$ or else it proves $(n)(\alpha)[\alpha(0) = n \supset (\alpha)(Ex)B(\bar{\alpha}x)]$. Put precisely we consider a property $P(\mathfrak{a})$ of formulae of the form $(\alpha)[\alpha \subset \mathfrak{a} \supset (Ex)B(\bar{\alpha}x)] : (\mathfrak{a})\{[B(\mathfrak{a}) \supset P(\mathfrak{a})]$

[25] Cf. the corresponding problem in classical analysis, for arithmetic $A(P, x) : P_A$ is defined explicitly; to *prove* 2.612 for arbitrary Q one needs full comprehension; but a *model* for both 2.612 and 2.613 needs only the one function quantifier form: for these axioms are satisfied by functions of hyperdegree $<O'$; this follows from [30].

[26] And P_A in turn only positively, that is, as restrictions of quantifiers $(x : P_A(x))$, $(Ex : P_A(x))$. As a figure of speech: an inductive definition tells you what is *in* P_A, not what is *not* in P_A. If there is something behind this idea, it probably has to be analysed by a different logic since, on 2.2, the meaning of $\sim P_A(x)$ is determined if $P_A(x)$ is determined.

[27] Or: "can" be put into this form; as any logical proof "can" be brought into the form described by certain formal rules.

$\& [(n)P(\mathfrak{a} * n) \cdot \supset P(\mathfrak{a})]\}$, that is, $(\mathfrak{a})\{B(\mathfrak{a}) \vee (n)P(\mathfrak{a} * n)] \supset P(\mathfrak{a})\}$ which is an inductive definition of P as in 2.611, with $B(x) \vee (n)P(x * n)$ for $A(P, x)$; if B is decidable, this is purely universal. Rephrased in terms of functionals (neighborhood functions): either $x(\alpha)$ is constant or, for a sequence x_n, $x(\alpha) = x_{\alpha(0)}(\alpha^*)$ where $(n)[\alpha^*(n) = \alpha(n + 1)]$. Define a species K of *functions* of finite sequences

2.621. $e \in K$: $e(<>) = 0 \ \& \ \{(\mathfrak{a})[e(\mathfrak{a}) = e(\langle 0 \rangle) \neq 0] \vee (n)(Eb)(\mathfrak{a})[e(n * \mathfrak{a}) = b(\mathfrak{a}) \ \& \ K(b)]\}$.

Thus we express the proposed condition on proofs by:

2.6211. $(\alpha)(Ex)A(\alpha, x) \leftrightarrow (Ee \in K)(\mathfrak{a}: e(\mathfrak{a}) \neq 0)(\alpha \subset \mathfrak{a})A[\alpha, e(\mathfrak{a}) - 1]$. Note that the quantifier α is pushed past (Ex), and appears only with a decidable restriction: $\subset \mathfrak{a}$.

2.622. *Second Elimination Result. In* $H_1{}^1$ *extended by all the additional axioms of* 2.5, 2.612 *and* 2.613 *for the particular* K *of* 2.621, *for every closed formula* A, *there is a formula* A' *not containing any Greek variables at all, such that:* $A \leftrightarrow A'$.

2.623. Some theorems of this theory (2.622). First, the converse to 2.6153; this follows from Brouwer's *bar theorem:* if a decidable (binary) relation R, not necessarily an *ordering*, is well founded in the sense $(\alpha)(Ex) \sim [\alpha(x + 1)R\alpha(x)]$ then proof by transfinite induction on R applies to arbitrary Q, that is, $(x)[(y \ R \ x)Q(y) \cdot \supset Q(x)] \supset Q(x)$. This in turn follows from

2.6231. If $R(\alpha, x)$ is monotone (not necessarily decidable), that is, $(\alpha)(x)[R(\alpha, x) \supset R(\alpha, x + 1)]$, and Q is an arbitrary property of finite sequences then

$$\{(\alpha)(Ex)R(\alpha, x) \ \& \ (\alpha)(x)[R(\alpha, x) \supset Q(\bar{\alpha}x)] \ \& \ (\mathfrak{a})[(n)Q(\mathfrak{a} * n) \cdot \supset Q(\mathfrak{a})]\} \supset (\mathfrak{a})Q(\mathfrak{a}).$$

And this is proved by induction on K. Conversely

2.6232. If Brouwer's bar theorem and the two axioms for K are assumed we get 2.6211.

2.6233. If K is (explicitly) defined in $H_1{}^1$ by: $e(<>) = 0 \ \& \ (\alpha)(Ex)(y)[e(\bar{\alpha}x * y) = e(\bar{\alpha}x) \neq 0]$, 2.621 and the corresponding principle of induction are theorems of: $H_1{}^1$ together with the bar theorem.—N.B. The familiar conversion of inductive into explicit definitions (footnote 25) is applicable in the intuitionistic case for a more restricted class of $A(P, x)$ in 2.613 than in the classical case (cf. footnote 24).

2.6234. Not only do the e in K define continuous functionals,

but uniform continuity can be proved on functions dominated by a (fan theorem), x being the modulus of continuity:

$$(a)(e \in K)(Ex)(b)(c)(\{(z \leq x)[b(z) = c(z)] \,\&\, (z)[b(z) \leq a(z) \,\&\, c(z) \leq a(z)]\} \supset [e(\bar{b}x) = e(\bar{c}x) \neq 0]).$$

2.6235. By essential use of the continuity axioms 2.5, the *bar theorem of type 1* (for relations R between type 1 objects) can be proved in 2.622, when formulated as follows:

$$(\alpha)(Ex) \sim [\lambda y \alpha(< x + 1, y >) R \lambda y \alpha(< x, y >)] \supset \{(\alpha)[(\beta R \alpha) Q(\beta) \supset Q(\alpha)] \supset (\alpha) Q(\alpha)\}.$$

2.624. *Refinements and Alternatives.* If one is interested in consistency: there is a finitist argument showing

2.6241. If A is theorem of 2.622, A' can be proved in H_1 extended by the axiom of choice and the two axioms for K. Consistency may be in question because 2.622 cannot be *extended* to a nonconstructive theory: it is inconsistent if the law of the excluded middle is added, while H_1 with K is not.

The latter system may still look stronger than it is because K is a set of *functions*. However,

2. 6242. H etc. may be reduced to H itself with the corresponding inductive definitions for *recursive* neighborhood functions. The application of the axiom of choice needed is in the form:

$$(n)(Ee)[(e \in K) \,\&\, A(n, e)] \supset (Ee_1)\{e_1 \in K \,\&\, (n) A[n, \lambda a e_1(n * a)]\};$$

this is supplied by the recursion theorem. Also

2.6243. H_1 etc. may be reduced to H^1 with the bar theorem. This latter system is of the same strength as that of footnote 25.

If one drops constructive functions altogether, and considers H^1_{ω} + axiom of choice + bar theorem + continuity, as in [8], there is an alternative route to H^1 + bar theorem itself. First apply the interpretation 2.321 to get a reduction to a quantifier free system (BR) (so-called bar recursion of lowest type 4.222 with c ranging over finite sequences of numbers), then, as in 2.512, define in H^1 neighborhood functions with *modified* C_τ^E, and prove the existential axioms for (BR) by means of the bar theorem.

2.6244. It may be mentioned that in Section VI of the privately circulated Stanford report on "Foundations of Analysis" there is an elegant application of 2.321 to the system H_1 with the axioms of choice and for K, in which two *basic* types (cf. 2.51) occur in the

resulting quantifier free system $T(K)$. This has the advantage over 2.6242 of giving simple closure conditions to be satisfied by models of the system considered, instead of using *all* recursive functions.

2.625. It would perhaps be interesting to develop a parallel theory in which definitional equality (2.35) between *constructive* functions is used. Then 2.621 is modified by use of combinators N (for numerical) and S (for shift): $(Nn)(<>) = 0$, $(Nn)(\mathfrak{a}) = n + 1$; $\{S(n, e)\}(<>) = 0$, $\{S(n, e)\}(\mathfrak{a}) = e(n * \mathfrak{a})$ and $(n)(Nn \in K_D)$ & $(e)\{(n)[S(n, e) \in K_D] \supset e \in K_D\}$, and to extend the analysis of [67], which provides a *decision function* for definitional equality (cf. 2.35);

Conjecture. The decision function for terms built up from variables of *bounded* type is itself definable [in the system $T(K_D)$ of 2.6244].

In this case, free choice sequences α^τ of *constructive* functions of higher type can be treated as ordinary free choice sequences chosen from a formally decidable spread (namely, the spread of numbers of *normal terms* of [67]).

2.63. *Discussion.* The results of 2.4–2.6 give a pretty full picture of the interreducibility between the theories described. This interreducibility is complete subject to the conjecture of 2.625; for then the *traditional* inductive definitions of constructive objects of higher type [purely positive $A(P, x^\tau)$ in 2.613] are first eliminated in favor of free choice sequences α^τ (cf. of 2.6233), and the latter reduced by 2.625. Inductive definitions of species P of free choice sequences are reduced by the *elimination results* to inductive definitions of species of constructive functions: $a \in P_C \leftrightarrow [S(a)$ & $(\alpha \in a)P(\alpha)]$.

The matter is open for monotone $A(P, x^\tau)$.

Note in passing that the objects known as absolutely free (better: lawless) choice sequences, mentioned on page 110 are not treated here. They are characterized by the condition that, from a certain point onward, no restrictions are to be made; that is, they are *given* by some initial segment, instead of arbitrary spreads. Therefore 2.521 is replaced by: $A(\alpha) \supset (E\mathfrak{a})[\alpha \subset \mathfrak{a}$ & $(\beta \in \mathfrak{a})A(\beta)]$, as in [47]. Though neither kind of choice sequence is a special case of the other, they are interreducible via generalized inductive definitions. The lawless sequences have only been used as examples in completeness proofs (cf. 2.74) and for pedagogic purposes as on page 110.

Concerning the mathematical content of the systems above:

though 2.622 shows weakness compared with full analysis, a good deal of impredicative analysis (cf. Section 3) can be developed in H^1 together with bar induction; cf. [41] to which should be added theorems of the Cantor-Bendixson type.

For a figure of speech expressing continuity, the analysis 2.6212 of the quantifier $(\alpha)(Ex)$ is completely satisfactory (while $(\alpha)(Ex)$ does not yield the continuity properties). Heuristically this could be expected: In nonconstructive analysis, we have of course the ordinary definition of continuity $\mathfrak{C}$ for mappings: $N^N \to N$ (N^N with the product topology); then it is a *theorem* that $\mathfrak{C}$ is the least class of mappings ϕ: constants $\epsilon\mathfrak{C}$ and if $\phi_0, \phi_1, \ldots \epsilon\mathfrak{C}$ so does $\lambda\alpha\phi_{\alpha(0)}(\alpha^*)$. Now, it so happens that if one starts with this as *definition*, most of the natural proofs simply do not use nonconstructive principles. As an assertion about intuitive proofs, 2.6241 shows that Brouwer's proposal is consistent with the remaining (accepted) axioms; here it may be added that for the usual *formal* systems the proposal is actually verified (cf. 3.341), where necessarily infinite proof trees are used (as stressed by Brouwer [20]).[28]

Note that *nothing has been established here on free choice sequences of proofs*, that is, on extensions of the theory 2.21 corresponding to 2.5.

2.7. *Church's Thesis.* Though the best proof theoretic results (for example, 2.6244) depend on the discovery of special definitional schemata, a general way of looking at the results 2.4–2.6 is this: for the axioms considered we have a *reduction* of higher types to arithmetic. The first nontrivial case concerns type 0^0 objects and is expressed in the thesis:

Every constructive number theoretic function has an equivalent definition by means of a certain kind of computation procedure.

The most familiar kind is: by means of recursion equations with a suitable equation calculus. These equations can be numbered, and, by [7], there is a primitive recursive relation $T(n, m, p)$ which holds if p is the code number of a computation from the equation n for the value $\{n\}(m)$ (of the function $\{n\}$ defined by n).

2.71. *The Meaning of Church's Thesis.* We begin with some distinctions.

2.711. The fundamental distinction of [34] between finitist (com-

[28] These are in any case much closer to the way proofs, for example in number theory, actually *present* themselves to us: the use of dots in proofs by induction is not an accident!

binatorial) and constructive *proof*, carries over to combinatorial and constructive *definability*, for example, of number theoretic functions. The definitions of number theoretic functions of 2.22, as intended, refer to abstract objects (higher types), and are only later reduced to arithmetic. More technically: if one understands the meaning of the symbols, the equations describe the functions uniquely, and it is then a theorem that the particular calculus used is complete.

The distinction between 2.712 *combinatorially definable* and 2.713 *finitist* function is then this: a combinatorial rule is given (that operates only on finite configurations), and an arbitrary constructive proof of the termination of the procedure is accepted in 2.712, but only a finitist one (cf. Section 3) in 2.713.

2.714. Most of the literature speaks of: effectively calculable, instead of: constructive functions. Probably 2.712 (and not the more general sense of constructive definability) was meant because the implicit aim was to provide instructions suitable for a moron or a mechanism. Perhaps a better word for 2.712 would be: mechanically definable. Not (as far as I know): physically definable. For, it is true (excepting collisions as in the 3-body problem, which introduce discontinuities) the theory of partial differential equations shows that the behavior of discretely described (finite) systems of classical *mechanics* is recursive. But this may not be so in the quantum theory, for example, of large molecules.[29]

2.715. The support for Church's thesis applies to 2.712, and not to the general notion of constructive definability. It consists above all in the analysis of machine-like behavior and in a number of closure conditions, for example, diagonalization; cf. [7]. It certainly does not consist in the so-called empirical support; namely the equivalence of different characterizations: what excludes the case of a *systematic* error? (Cf. the overwhelming empirical support from ordinary mathematics for: if an arithmetic identity is provable at all, it is provable in classical first order arithmetic; they all overlook the principle involved in, for example, consistency proofs.)

[29] In other words: is this theory a theory in the sense of Dirac [4], that is, one which permits mechanical approximations? Consider (in the language of statistical mechanics) the "cooperative phenomenon" (like boiling) of the mathematical community and its asymptotic behavior with respect to arithmetic problems. It looks pretty stable. Do we have a better theory of what answers it will give to such problems than: according to arithmetic truth? *This* theory is certainly not recursive.

2.72. The formulation of Church's thesis. We express the thesis in (extensions of) H_1 by

$$(a)(En)(m)[(Ep)T(n, m, p) \,\&\, \{n\}(m) = a(m)].$$

2.721. If no variables for constructive functions are available, for example, in H^1, we use the schema, for *all* A not containing *free* Greek letters:

$$(x)(Ey)A(x, y) \supset (En)(m)\{(Ep)T(n, m, p) \,\&\, (x)A[x, \{n\}(x)]\}.$$

This follows in our systems from 2.72 since $(x)(Ey)A(x, y) \supset (Ea)(x)A[x, a(x)]$.

2.722. By 2.714, 2.72 *asserts* a little more than Church's thesis since the axioms of H_1 are not especially evident if the a are interpreted as constructive functions in the sense 2.712. Nevertheless

2.723. *The systems of 2.4–2.6 remain consistent if 2.72, respectively 2.721, are added.* The proof for systems not containing variables for free choice sequences is straightforward. For other systems one first applies the elimination results.

2.7231. As is to be expected from the intended meaning underlying [8], the consistency of 2.721 for the system there discussed (essentially H^1 with continuity and the bar theorem) is not *shown* in [8]. However, by 2.6, 2.721 can be consistently added. Note that [8] establishes the consistency of the *rule:* if the premise of 2.721 is proved, add the conclusion.

Evidently 2.72 allows a reduction to an arithmetic system. More generally, 2.4 constitutes a reduction of certain systems for all finite types by means of

2.73. Extensions of Church's thesis to all finite types. Here we go back to the basic notions of higher type in 2.331 and distinguish between

2.731. Applying Church's thesis simultaneously at all types: then 2.331 (i) (with extensionality) become the so-called effective operations of [46], and in 2.331 (ii) all neighborhood functions considered are recursive. By [46] (4.21) the same classes of functionals are obtained.[30]

2.732. Applying Church's thesis only at highest type, that is, at type ξ ($= 0^\tau$ say), one requires a recursive computation rule, but the objects of type τ are only assumed to be constructive, not themselves recursive. In case 2.331 (ii), where continuity is assumed,

[30] Note, however, that the equivalence proof in [46] is nonconstructive, cf. *negative* continuity of [20].

the results are fairly complete. One considers free choice sequences satisfying the conditions C_τ for neighborhood functions of type τ; since these are formulated in H^1, one has a reduction to intuitionistic analysis. The most interesting conclusions concern bar recursion (cf. 4.3 below). The case 2.331 (i) is more delicate. The most interesting extension of Church's thesis to this case is in [44], but unfortunately it contains no analysis of the axioms for functions of higher type used in the proofs of [44]: so it is not yet known if any nontrivial results stated there are valid on the basis of evident properties of *constructive* functions of finite type.

2.7321. At low types the properties of the functionals of [44], that is, defined by S1–S9, are satisfactory; for example, for type 2, $= 0^{0^*}$, if such a functional can be proved to be defined for all type 0^0 objects (not assuming Church's thesis), it is defined for all free choice sequences. For, we have no axioms which exclude the possibility that there is a rule for each free choice sequence. But if one proves that a type 0^2 object is defined for all type 2 objects, one cannot make use of the continuity properties of type 2 objects. This is satisfactory because we have no axioms which ensure that all constructive functions of type 2 *are* continuous.

2.7322. Regarded as existential axioms, the schemata S1–S9 of [44] are weak (without additional axioms about functions of higher type) because, for example, the effective operations of 2.731 are closed under these schemata, by [46], 4.32.

2.74. Formally speaking, Church's thesis plays a somewhat similar role in present-day intuitionistic mathematics as the ramified hierarchy in set theory. Not only is it consistent with the known axioms, but it can also be used to show the *formal* character of interesting open questions; cf. page 109. One important result (mentioned in 2.732) is given in 4.32. Another is this:

2.741. The rules of intuitionistic predicate logic cannot be proved complete by any methods consistent with Church's thesis. For, on Church's thesis, there is a primitive recursive binary tree (spread) B, such that $(a)_B(Ex)A(\bar{a}x)$, where a_B ranges over 0, 1 sequences, but $\sim(a)_B(Ex)A(\bar{a}x)$, that is, by 2.524, $\sim[(a)_B \sim \sim(Ex)A(\bar{a}x) \supset (a)_B(Ex)A(\bar{a}x)]$. By [48] this implies incompleteness.

2.7411. *A Distinction.* In terms of intuitionistic primitive notions, there are two meanings of completeness of a set of formal rules: (i) every *proof* for an assertion (in the given language) is

(among those represented by) some formal derivation; (ii) every assertion which is provable at all *also* has a formal derivation. Only (ii) can be expected; for summary, cf. [48], and the above incompleteness applies of course to (ii). It may be remarked that even in the propositional case no completeness results for non-negative formulae are known without the use of absolutely lawless sequences of [47]. It would be interesting to study this matter.

2.75. *Is Church's Thesis Valid?* More precisely, is 2.72 valid, where a ranges over all constructive number theoretic functions, and not only combinatorially definable ones in the sense of 2.712, since the latter are not basic for intuitionistic mathematics, as expounded here. Evidently there is no reason why the question should not be decidable by means of evident axioms about constructive functions. After all, work was needed to establish 2.723! It might have turned out to be false. The discovery of axioms about constructions which are inconsistent with Church's thesis is certainly one of the really important open problems.

Among the subjects discussed above, the most likely candidate is 2.613 for arbitrary *monotone* A. In view of the *conjecture* in 2.625, the axioms for higher type *so far* formulated are probably consistent with Church's thesis.

2.8. *Constructive Analogue to the Cumulative Theory of Types.* The following set of axioms is offered as a formulation. It may be redundant, that is, some of the functors introduced may be obtainable from the other axioms. More important, some obvious principle may have been overlooked; all that is done is to indicate a *kind* of theory. It is the natural analogue of the theory of types in 1.1.

Variables $a, b, \ldots$ over constructions; constants 0, 1; functors T (the type of a construction), D, D_1, D_2 (for forming sequences), E (decision function for $=$), application, and the relation symbols $=$, $<$ (partial order for constructions which are ordinals). The monadic predicate symbol O is defined by: $O(a) \leftrightarrow (0 < a \vee a = 0)$.

Intuitionistic Predicate Logic. Term formation as in 2.2. Instead of extensionality: $E(a, b) = 0 \vee E(a, b) = 1$, $E(a, b) = 0 \leftrightarrow a = b$. Inductive *definition* of $<$: instead of the least element principle, the (classically equivalent) contrapositive, and *restricted* trichotomy $(b < a \,\&\, c < a) \supset (b = c \vee b < c \vee c < b)$; however, these are not axioms but consequences of the following generalized

inductive definition:

$$0 < D(0,1),\ (a < b\ \&\ b < c) \supset a < c,\ a < D(a,1) \qquad \text{(successor)}$$
$$(0 < b < c < a)[d(b) < d(c)] \supset (b < a)[d(b) < D(d,a)] \qquad \text{(supremum)}.$$

Corresponding *principle of proof* for arbitrary binary relations $Q(a, b)$.

Type structure:

$$O(Ta);\ T0 = 0,\ T1 = 0,\ Ta < T[D(a,b)] = T[D(b,a)]$$
(these may be insufficient $O(a) \supset Ta = a$?).

Comprehension: For each term $t[x]$,

$$(Eb)(c)(Tc < Ta \supset b(c) = t[c]).$$

Type limitation: $a(c) \neq 0 \supset Tc < Ta$
Axiom of choice: for each formula $A(x, y)$

$$(b)[Tb < Ta \supset (Ec)A(b,c)] \supset (Ec)(b)(Tb < Ta \supset A[b, c(b)])$$

Existence of an infinite ordinal:

$$0 < \omega, \qquad (a)(a < \omega \supset \langle a, 1\rangle < \omega).$$

2.811. About the axiom of choice above: construe all c-cum-type-a as functions of both b *and* the proof of $Tb < a$. In the recursive case (for $a \in O$ of 2.6141) the primitive recursive functions obtained by the recursion theorem correspond to *functors* in the abstract theory which are not subject to type restriction.

2.812. In Brouwer's theory of ordinals the supremum axiom is assumed only for $a = \omega$.

2.813. I have not checked whether the principles above allow one to derive ordinary induction for a suitably *defined* ω; if not, one must regard the existential axioms for ω as an inductive definition. From this one can then develop enough elementary arithmetic to state Church's thesis.

2.82. *Is Church's Thesis Consistent with This Theory?* Two

comments. On page 121: in view of the elimination of free choice sequences, the theory *may* be as adequate for present-day intuitionistic mathematics as the cumulative type theory for classical mathematics, but this is not yet known. Also, to make it really constructively evident, a more elaborate analysis of transfinite types, as in [67], may be necessary. Second, in the formulation above (use of E) all terms must be assumed to be fully defined [with the convention: $a(c) = 0$ if $a(c)$ is not defined (for a, c, as *given*), to satisfy type limitation; in contrast to 2.15]. Thus, for instance, there is no direct contact with so-called higher number classes [58] where *partial* recursive increasing sequences are assumed to have a supremum. Most of the literature on them is nonconstructive; the discovery of laws for such partial operations (correct on the intuitionistic conception) may lead to quite a new view of the subject; cf. footnote 16. (They may be those of current intuitionistic logic itself where $=$ is decidable only at type 0.)

3. PROOF THEORY

The reader has probably a clear idea of what a *formal system* is (axioms and inference rules); for a systematic treatment cf. [13]. But he has probably never *used* one: for instance, to convince himself that (a statement) A is not a formal consequence of B, he generally finds a model (= example) of B in which A is false, and rightly assumes that if the formal rules are sensible, A cannot be formally derived from B; cf. page 2.121. In contrast, in the present section rules will play a central role; cf. 3.24.

Notation: Throughout this section we use symbols $\neg$, $\wedge$, $\vee$, $\bigwedge$, $\bigvee$ for the "logical constants" of (possibly uninterpreted) formal systems, the notation of Section 1 for truth functions, and of Section 2 for intuitionistically interpreted formal systems.

As explained in the introduction the basic formalist contention is that the inherent structure of intuitive proofs can be represented in terms of formal manipulations of symbols. As far as the set of *provable* statements of elementary logic is concerned, suitable rules were given by Frege in the last century, the possibility of obtaining all intuitively valid logic rules was first verified empirically, and much later given a theoretical justification in the completeness theorem (1.835). The mere generation of the theorems of a branch

of mathematics by means of formal rules may be called *crude* formalization.[31]

The mere existence of such rules does not explain (or justify) them; this can be seen as follows. Suppose we have a formal derivation of the sequence of seven symbols: $a + b = b + a$. Why are we to expect that addition of numerals (by correct application of the usual computation rules) is independent of order? In other words: how is the formal derivation connected with the application we make of it, even, as here, within mathematics?[32]

It is not immediately apparent how this question is to be significantly formulated in a formalist frame work. As mentioned on page 97 there is the possibility of a conflict. Old-fashioned formalism (for a revival and forceful exposition, cf. Bourbaki) rejects the question—at least as being capable of primarily theoretical treatment—and proposes to treat it on an empirical (statistical) basis.

Trivially, as long as all formal rules are considered "theoretically equal," one cannot formulate questions of their *evidence*.[33] Hilbert formulated a coherent proposal for treating such questions which is described below. It can be judged, of course, on its merits; however, if one has been weaned on old-fashioned formalism, it is helpful to consider first the *weakness* of the latter position. First, in practice, the reliability of mathematical principles is not studied

[31] The conviction, probably, is not merely that such rules happen to generate the provable statements in a particular domain of mathematics, but that (despite appearances, footnote 28) this is really all that goes on. Just as in the early applications of the molecular theory in chemistry one did not merely mean that the particular integral ratios in chemical reactions happen to be *formally* explained by an atomic hypothesis, but that there were such things as atoms. As observed above, the formalist contention does not get much support from naive observation; *officially*, it was proposed on general philosophical grounds for its freedom from metaphysical (ontological) assumptions of the existence of abstract objects such as *sets* (Section 1), *mental acts* (Section 2). But, probably, its *attraction* derives at least partly from this: long before electronic computers were thought of, one could see more or less how behavior according to such formal rules could be realized by a *mechanism*, that is an old fashioned mechanism in the sense of a Turing-machine (cf. footnote 30). Thus it gave the best hope of a mechanistic theory of reasoning.

[32] Note for reference: the question arises even for small numbers, say $a, b < 20$, if we have not tried out all 190 cases.

[33] As one says: the theory lacks the multiplicity of the phenomena to be explained, cf. footnote 3.

statistically; if it were, it would be essential, before accepting the principle of the least upper bound, for example, to find out whether the conclusions one has actually drawn from it do not use very special instances only (cf. page 118). Second, the matter is problematic even if only short expressions are considered, cf. footnote 32. Third, more important, no attempt has been made even to state the principles to be used in such a statistical study. One suspects that this would lead back to just the kind of question on evidence which is being rejected. Incidentally, a serious analysis of the matter should be very interesting.

3.1. *Hilbert's Program.* The remark about the commutative law (or, for that matter, about statistical inference) shows that the crude formalization already described has to be supplemented by an analysis of principles of *evidence;* for positive results these will be codified in particular formal systems whose formal derivations (are recognized as representing proofs which) have this kind of evidence (cf. page 122). Then the problem is: *Suppose we are given a codification (in the crude sense above) of a branch of current mathematics; to reduce it to proofs of the particular kind considered.*

3.11. What form should the reduction take? Hilbert's basic considerations were (more or less) these:

3.111. The type of reasoning which constitutes the most elementary analysis of such things as the commutative law, is of the kind expounded systematically by Grassmann, and codified, for example, in primitive recursive arithmetic; it refers to purely concrete situations, in our particular case to the very rules for computing sums. This is a type of reasoning used in all theoretical work, i.e. constitutes a minimal presupposition. So if the reduction, in whatever form it is formulated, only requires this kind of reasoning, it has the maximal kind of evidence that can be reasonably demanded.

3.112. The property of being a formal derivation in a formal system, or the relation between premise and conclusion in an inference rule, is of the same arithmetic kind (use any of the familiar codings of finite strings of symbols by means of numbers).[34] In consequence we have an arithmetic relation $\mathrm{Prov}_F(n, \ulcorner \mathbf{A} \urcorner)$: n is a formal derivation by the rules F of the formula **A** (and $\ulcorner \mathbf{A} \urcorner$ is a term whose value is the number of **A**).

[34] Or, conversely, represent numbers by strings of symbols! (Concatenation theory.)

3.113. *Basic Point.* It is clear that, for the *intended* interpretation of the formal languages in Sections 1 and 2, most formulae do not have purely combinatorial significance at all, because they refer to abstract situations, for example, if they have quantifiers over sets, or (implied quantifiers in) implications in Section 2. These formulae Hilbert called *ideal* elements of the language used to round it out. But certain relations are intended to *correspond* to combinatorial ones; precisely, for an elementary relation $P(n)$, we may have a set theoretic or intuitionistic relation represented formally by $\mathbf{P}(x)$, the term $\mathbf{S}^n\mathbf{0}$ denoting the nth numeral, and we have (elementary) proofs of:

3.1131. $$P(n) \Rightarrow \mathrm{Prov}_F\,[\pi_P(n), \ulcorner \mathbf{P}(\mathbf{S}^n\mathbf{0}) \urcorner]$$

3.1132. $$\overline{P(n)} \Rightarrow \mathrm{Prov}_F\,[\pi_{\bar{P}}(n), \ulcorner \neg \mathbf{P}(\mathbf{S}^n\mathbf{0}) \urcorner]$$

where $\neg$ is the (formal) *negation* symbol of F, π_P, $\pi_{\bar{P}}$ (elementary) functions which essentially describe how the intended computation of $P(n)$ can be mimicked in the abstract theory. Formalization before Hilbert had actually provided such proofs (= nothing else but the development of arithmetic in set theory, and so on). Now *reducing proofs of assertions of this combinatorially significant kind to combinatorial ones* is expressed (for the formal variable x of F) by:

3.1133. $$\mathrm{Prov}_F\,[m, \ulcorner \mathbf{P}(x) \urcorner] \Rightarrow P(n)$$

This statement now is formulated entirely by use of purely combinatorial terms, and therefore there is at least a possibility of proving 3.1133 on the minimal assumptions mentioned in 3.111.

Hilbert also emphasized the amusing fact that, granted 3.1131 and 3.1132, 3.1133 is an immediate corollary of the *consistency statement* (with variable formula $\mathbf{A}$), variables m and n over (numbers of) formal proofs:

$$\mathrm{Prov}_F(m, \ulcorner \mathbf{A} \urcorner) \Rightarrow \overline{\mathrm{Prov}_F(p, \ulcorner \neg \mathbf{A} \urcorner)}.$$

3.12. *Some Pragmatic Comments.* (a) It is clear that there might be ambiguity in what exactly constitutes an elementary (consistency) proof. If one really expects to find one, one does not split hairs beforehand, but gives a precise formulation only after one has found such a proof. The same applies, for example, if one asks for a *practical* or *nice* solution of any other problem. Only when it comes to negative results is an analysis needed. (b) The

same applies to possible ambiguities in the arithmetic representation of the combinatorial relation: $\mathrm{Prov}_F(n, \ulcorner \mathbf{A} \urcorner)$.[35] (c) Of course, a proof of consistency is a minimum requirement for the kind of reduction required; but the central role of consistency depends on the particular kind of evidence here studied where only purely combinatorial assertions are "real." In a proof theory which uses less elementary methods (and therefore is useful for a reduction to a less elementary kind of evidence), it would be quite unreasonable to give a special place to the consistency problem. (d) In the light of past research the consistency problem is recognized as technically difficult. If one really holds the mechanistic conception, one will of course regard it as important, but not necessarily difficult. Just as, when one believes the equations of classical mechanics to be true, one will regard as important the verification that initial positions and velocities determine future behavior (subject to given forces), but not necessarily difficult.

3.2. *Basic Facts about Formal Systems* (limitations and general considerations on the choice of systems). Throughout this section we use the following

3.21. *Distinction.* Instead of considering formal systems, that is, a collection of rules, one sometimes considers equivalence classes, where systems with the same set of theorems are identified. This is often appropriate for questions which refer just to the set of theorems, for example, whether it is recursive [16] or saturated, that is, for each (closed) formula $\mathbf{A}$, either $\mathbf{A}$ or $\neg\mathbf{A}$ is a theorem. In this case, the natural requirements 3.1131 and 3.1132 for *representability* of a property $P(n)$ may be weakened: $P(n) \Rightarrow \vdash_F \mathbf{P}(\mathbf{S}^n\mathbf{0})$ and $\overline{P(n)} \Rightarrow \vdash_F \neg\mathbf{P}(\mathbf{S}^n\mathbf{0})$ are only required to be true for each n (in the classical case; and constructively, not necessarily elementarily, provable in the intuitionistic case). Much of this work uses little more than the recursive enumerability of the set of theorems of the formal systems considered, and is therefore best treated as part of recursion theory, cf. [13]. (However, there are amusing differences of detail [25].) For finitistic refinements, cf. 3.3042.

3.22. Rules have to be considered when one is interested in

[35] This ambiguity is liable to be overlooked if one assumed that, for two combinatorially defined $P(n)$, $P'(n)$, either one can find n_0, $\overline{P(n_0)} \Leftrightarrow P'(n_0)$, or else there is an elementary proof of: $P(n) \Leftrightarrow P'(n)$.

questions of evidence since the equivalence of different sets of rules may simply not be evident. Obviously, if $\mathbf{Prov}_F\ (n, m)$ is one representation of the proof relation for one of the usual systems, and $\mathbf{Prop}\ (m, m')$ expresses that m is propositionally equivalent to the negation of m' then

3.221.

$$\mathbf{Prov}_F\ (n, m) \wedge \bigwedge p < n \bigwedge m'[\mathbf{Prop}\ (m, m') \rightarrow \neg\, \mathbf{Prov}_F\ (p, m')]$$

is another representation; for 3.221, consistency is trivial since derivations leading to inconsistencies are forbidden. (N.B.: if **A** and $\mathbf{A} \rightarrow \mathbf{B}$ are provable in sense 3.221, not necessarily **B**!)

3.222. Elimination of ambiguities (cf. footnote 35). If a precise analysis of formal systems is given [13], ultimately: a *finite* set of rules (for generating possibly infinitely many axioms), the *canonical representation* (c.r.) in F of the rules is determined up to formal equivalence in F by this: the rules have the form of an inductive definition (2.612, 2.613 with quantifier free positive A) of the property: **A** is formally derivable; then the formula $\mathbf{P}(x)$ of F *represents* this property *canonically* if 2.612, 2.613 are derivable in F for **P** in place of P. Then with a minimum of arithmetic in F, for two c.r. **P**, **P'** we have $\vdash_F \bigwedge x[\mathbf{P}(x) \leftrightarrow \mathbf{P}'(x)]$.

3.2221. To apply to quantifier free systems, minor modifications must be made.

3.2222. It should be remarked that though a theory of c.r. is clearly possible it has never been worked out in full in the literature: an elegant exposition would do much to make the elements of proof theory more systematic.

3.2223. By 3.12, these general considerations are not specially needed for positive results, to be given in 3.3. But they help to formulate the

3.23. *Fundamental Incompleteness Result* [32]. Suppose F is a formal system, consistent, and a certain minimum of arithmetic can be developed in F (for precise conditions, cf. [13]; in particular, the proof relation itself is represented in the sense of 3.21), then

3.231. There is an elementary, even primitive recursive property, P represented by $\mathbf{P}(x)$ such that neither $\vdash_F \bigwedge x\mathbf{P}(x)$ nor $\vdash_F \neg \bigwedge x\mathbf{P}(x)$; in fact, for each n, $P(n)$ is true (because if $\vdash_F \neg\mathbf{P}(\mathbf{S}^n\mathbf{0})$ for some n, $\vdash_F \neg \bigwedge x\mathbf{P}(x)$ in the F considered).

3.2311. While there is an *axiomatic* characterization of the notion of natural number (relative to the notion of set) in the sense of

1.32, there is no *deductive* characterization (by means of a formal system) in the sense of yielding precisely the true statements in the usual notation. (Axiomatic characterizations in 1.32 are so-called *second* order axiomatic systems.)

3.232. Under essentially the same conditions as in 3.231, there is a formula **A** such that, if **Prov** (n, m) represents the proof relation, not $\vdash_F \wedge x[\mathbf{Prov}(x, \ulcorner \mathbf{A} \urcorner) \rightarrow \mathbf{A}]$.

3.2321. For the usual proofs, **A** is $\wedge x \mathrm{P}(x)$ of 3.231 because it expresses its own nonderivability: $\mathbf{A} \leftrightarrow \wedge x \rightarrow \mathbf{Prov}\ (x, \ulcorner \mathbf{A} \urcorner)$.

3.233. Under essential use of provability of closure under detachment (cf. 3.221) and stronger representability conditions [52], but certainly for c.r. **Prov**, $\vdash_F \wedge x[\mathbf{Prov}(x, \ulcorner \mathbf{A} \urcorner) \rightarrow \mathbf{A}]$ if and only if $\vdash_F \mathbf{A}$.

3.2331. *Cor.* Taking $\mathbf{0} = \mathbf{1}$ for **A**, the canonical expression for the consistency of F cannot be proved in F.

3.231 is the natural formulation of inadequacy of formal systems for the notion of *arithmetic truth;* but it is not sufficient for inadequacy with respect to intuitive provability because one has no reason to suppose that, for each combinatorial P, $P(n)$ with free variable n is intuitively provable if it is true for each n.

3.234. *Discussion of* 3.232 *for the Mechanistic Conception.* We cannot construct a formal system with the following two properties: (a) all elementary proofs are represented in it; (b) its formal consistency is provable by elementary means. But both would be needed to support the mechanistic conception in its *full* sense: (a) is needed to say that a formal system is adequate for the *representation* of all elementary proofs; and (b) for the *reduction* to elementary evidence. But 3.232 does not exclude a practical validity for the mechanistic conception with a kind of transcendental singularity: (a) is false, but for all axioms for which we believe to have an abstract interpretation, (b) holds.

The general feeling has long been that the opposite is true; namely that classical arithmetic, and certainly classical analysis, satisfies (a). This requires a precise characterization of elementary proofs. 3.4 and 3.5 contain some relevant work and 3.6 open problems.

3.24. *Technical Considerations about Formal Systems* (rules versus axioms). The reader is probably most familiar with certain kinds of formal systems, namely so-called axiomatic theories in standard formalization [16] consisting of axioms and *valid* rules of inference, that is, schemata for which we recognize that, if the premises are

true so are the conclusions. This type of formalization is suitable when one wants to make as evident as possible that all theorems of the system are valid (for the intended interpretation). A totally different consideration is needed if one wants to characterize some such notion as elementary or combinatorial proof; specifically, if one reviews

3.241. Formalization in the light of the incompleteness theorem 3.2331. Formal systems (having a recursively enumerable set of theorems) are not suitable for characterizing the notion of arithmetic truth. If they are to be used for characterizing some (informal) notion of *proof*, it is essential *that each formal theorem represent an (informally) provable result, but that this fact be not recognizable by means of such an informal proof*. What this requires is just the opposite to the more familiar aim of formalization: we want rules whose correctness, for the notion considered, depends essentially on the fact that the premises have been *derived in a particular way*. To play safe (for avoiding conflict with 3.2331) one may have to use symbols which are not interpreted for the informal notion considered; for example, in 3.4 we shall use a symbol that is a kind of mnemonic.

3.2411. Naturally, the use of rules rather than axioms recommends itself also from different points of view, for example, [31]. And, purely pragmatically, it is clear that by using simple axioms and many rules of a simple structure (instead of few complicated axioms) one may make it evident that many interesting properties of the premises are inherited in the conclusions, and so, by induction, one shows that all formal theorems have these properties. This is often more useful for independence proofs than the construction of models (cf. 3.3232).

3.242. What rules should one study? Practically all results below are stated for intuitionistic systems. First, it turns out (3.31), the proof theoretic relations for the corresponding classical systems are read off simply by considering negative formulae (2.33) (while the converse is not true; thus the monadic intuitionistic predicate calculus is undecidable while the classical one is decidable [49]). So, the true generality of the results is hidden if one formulates them for the classical case only. It would be interesting to have a quite abstract formulation for general formal systems which isolates the structural properties of the (intuitionistic) rules actually relevant to the main results. But this has not yet been done.

Second, in the classical case one often gets substantially simpler semantic arguments for the results, at least if they are formulated in the traditional manner. In other words, by using the semantic validity of the rules in the sense of 1.831. The use of (the corresponding) 2.311 has not yet led to similar simplifications, and so, in the intuitionistic case, proof theoretic methods are not artificial.

3.243. A matter of taste: even when our primary interest is in getting particularly elementary *proofs* of a result, we try to reformulate the result so as to *force* the proof to be elementary. This is illustrated by the following stages of the consistency problem (for precise definitions, cf. 3.332). For any of the ordinary systems of Section 1 or 2, the problem is trivial without restrictions on the method of the consistency proof because the systems are valid under the intended interpretation. The big step is to ask another question which implies consistency in an *elementary* way: can the usual rules be replaced by so-called cut free rules? This is not immediate, but almost so (in the classical case): one shows the completeness of the cut free rules and concludes elimination from the validity of the usual rules. But if we ask: on what measure of complexity of a proof with cuts does the length of the corresponding cut free proof depend? then the *natural* solution is by elementary means. Such "computational" reformulations have been useful in the past for locating errors in syntactic arguments, for example, [17] by [6], page 123, or [38] by [24]. Cf. also 3.351 below.

3.2431. It is familiar from ordinary axiomatic mathematics that sharpened formulations may force proofs to be more elementary: a first order statement about rationals may be true but we have no idea how to prove it; if the same statement is true for all real fields, it is provable by first order predicate logic (from the axioms for formally real fields).

3.3. *Technical Proof Theory.* Most of the early work in proof theory was done in support of Hilbert's program and formulated accordingly. Now, first, Hilbert's finitist program can be carried out only for a limited range of mathematics; if this is to be extended at all, one must consider a reduction to (significant) principles of evidence other than finitist, and by 3.12(c) and 3.2 the consistency problem has to be widened. Second, extant proof theory actually establishes more than consistency. Naively, it might be supposed that each piece of proof theory has to be considered "on its merits."

But it turns out that significant results are conveniently summarized in terms of the notion of *interpretation* (of a system F in F_0). The syntactic relations will be established in a (partial) codification F_1 of interpreted proofs, in which the proof relations for F, F_0 have the c.r. $\mathrm{Prov}(n, m)$, $\mathrm{Prov}_0(n, m)$ respectively. For clarity, bold-face letters are used for expressions of (possibly) uninterpreted formal systems, and when the context makes the meaning clear, formulae are identified with their numbers and corners dropped. (a) In contrast to the consistency problem, one may consider a class $\mathfrak{A}$ of formulae in F other than c.r. of combinatorial properties. (b) If then $\mathfrak{A}$ contains quantified formulae and F_0 does not, simple examples suggest that with each $\mathbf{A}$ of F one should associate a disjunction $\mathbf{A}_1 \vee \mathbf{A}_2 \vee \cdots$ in F_0, for example for $\mathbf{A} = \forall x \mathbf{B}(x)$, $\mathbf{B}(\mathbf{0}) \vee \mathbf{B}(\mathbf{S0}) \vee \cdots$. (c) Finally, the interpretation should apply not only to proofs in F itself, but to the use of F for derivations from additional axioms ($\mathbf{A} \vdash \mathbf{B}:\mathbf{B}$ is derived in F from $\mathbf{A}$). Other properties of the notion are mentioned after the definition:

An F_1 interpretation in F_0 of a class $\mathfrak{A}$ of formulae in F is determined by function terms μ, π, σ (of F_1) with the following properties (n is a variable over formal proofs of F; $\mathbf{A}$, $\mathbf{B}$ over formulae of $\mathfrak{A}$).

3.301. For c.r. $\mathbf{A}$ of combinatorial statements, $\mathbf{A} \in \mathfrak{A}$ and also, for each n, $\mu(n, \mathbf{A})$ is the corresponding c.r. in F_0.

3.302. $\mathrm{Prov}\,(n, \mathbf{A} \vdash \mathbf{B}) \Rightarrow \mathrm{Prov}_0\,[\pi(n, m), \mu(m, \mathbf{A}) \vdash_0 \mu\{\sigma(n, m), \mathbf{B}\}]$ can be proved in F_1: taking for $\mathbf{A}(\mathbf{B})$ a true (false) combinatorial formula, we have from 3.301 and 3.302

3.3021. $\quad \vdash_1 \mathrm{Prov}\,(n, \mathbf{B}) \Rightarrow \mathrm{Prov}_0\,[\pi'(n), \mu\{\sigma'(n), \mathbf{B}\}]$

3.3022. $\quad \mathrm{Consistency}\ F_0 \vdash_1 \mathrm{Consistency}\ F$

Note. *Ad(a)*: 3.3022 shows that we really have a widening of the consistency problem. *Ad(b)*: $\mu(i, \mathbf{A})$ is $\mathbf{A}_i$. *Ad(c)*: 3.302 expresses that $\mathbf{B}_1 \vee \mathbf{B}_2 \vee \cdots$ follows in F_0 from each $\mathbf{A}_i$, that is, from $\mathbf{A}_1 \vee \mathbf{A}_2 \vee \cdots$.

3.303. If $F_0 \subset F$, an F_1 interpretation is called *proper* or *discriminating* if it satisfies in addition to 3.301 and 3.302 for $\mathbf{A} \in \mathfrak{A}$

$$\vdash_1 \mathrm{Prov}\,[\rho(n, \mathbf{A}),\ \mu(n, \mathbf{A}) \vdash \mathbf{A}],$$

that is, $\mathbf{A}$ is no stronger than $\mathbf{A}_1 \vee \mathbf{A}_2 \vee \cdots$ with respect to deduction in F. This permits a precise formulation of the relative subtlety of two formal systems, cf. 3.3231.

3.304. *Simplifications.* Suppose all the nonelementary *definitions* of the syntactic argument have been "built into F_0"; precisely, suppose π, μ, σ are quite elementary, for example, primitive recursive (PR). Then, if (one can prove in PR:) PR $\subset F_1 \subset F_0$ and if there is F_1 interpretation then there is also a PR interpretation, that is, 3.302 (which, by hypothesis, is *stated* in PR) can be proved in PR.

For, first generally (3.113), for $A \in$ PR, if $\vdash_1 A$ then: (Consistency F_1) $\vdash_{\text{PR}} A$; second, from an inconsistency in F_1 (and hence in F_0 since $F_1 \subset F_0$) we get a PR description of a formal proof in F_0 of *any* formula of F_0, in particular of $\mu(m, \mathbf{B})$. By combining the steps we have the result.

3.3041. If F_1 contains quantifiers, 3.302 simplifies to

$$\text{Prov}\,(n, \mathbf{A} \vdash \mathbf{B}) \supset (m)(Eq)(Ep)\ \text{Prov}_0\,[p, \mu(m, \mathbf{A}) \vdash_0 \mu(q, \mathbf{B})]$$

(where the notation of Section 2 is chosen because this is the most important case in applications).

3.3042. A formal instance of our interpretations is this: both F and F_0 contain the whole of classical logic (and therefore of no interest in the *present* context), but not only the consequence relation (3.302), but all logical connectives are required to be preserved by the mapping μ. These are the interpretations in the special sense of [16]. Though, by [13], the additional requirements are useless for the applications of [16], amusing results have been discovered for this notion in [26].

But the most interesting simplification occurs if we have

3.305. *Reflexiveness of F_1 over F_0 with respect to $\mathfrak{a}_0$*, that is, for each $\mathbf{A} \in \mathfrak{a}_0$ (possibly containing the free variable m) there is a corresponding (interpreted) formula A of F_1 and

$$\vdash_1 \text{Prov}_0\,(n, \ulcorner \mathbf{A}[\mathbf{S}^m\mathbf{0}] \urcorner) \supset A[m].$$

It is understood that $\ulcorner \mathbf{A}[\mathbf{S}^m\mathbf{0}] \urcorner$ is a term of F_1 (with variable m) which is the c.r. representation of the function giving the number of the formula obtained by replacing $\mathbf{x}$ in $\mathbf{A}[\mathbf{x}]$ by the mth numeral.

3.306. Suppose each $\mu(n, \mathbf{B})$ is an instance of a quantifier-free *schema* $B_0(s, t)$, that is, obtained by replacing s by a constant (functional) term of F_0. Then 3.3021 reduces to: $\vdash_1$ Prov $(n, \mathbf{B}) \supset (Es)(t)B_0(s, t)$, which is the form of the interpretation of logical constants of [34]; cf. 2.321 and used throughout Section 2. It satisfies 3.302.

3.31. The *interpretation problem* (3.3), like the consistency prob-

lem, refers to a, possibly proper, subclass $\mathfrak{A}$ of formulae in F, but proofs in F of $\mathbf{A}(\mathbf{A} \in \mathfrak{A})$ may contain formulae not in $\mathfrak{A}$. So even if we have an interpretation in mind, an inductive proof of 3.301–3.302 may be difficult. There are *two methods:*

Either to extend the interpretation to all formulae of F *or* to reformulate the rules of F so that in the new system F' a proof of $\mathbf{A}(\mathbf{A} \in \mathfrak{A})$ consists only of formulae in $\mathfrak{A}$ itself.

A very simple, but particularly useful, example of the former is this. F is classical (predicate logic or) arithmetic, and F_0 intuitionistic arithmetic. One takes for $\mu(\mathbf{A})$ the negative version of $\mathbf{A}$ (2.3322) and verifies that the rules of F are *derived* rules for the negative fragment of F_0. 3.302 follows, because, for closed $\mathbf{A}$, and the deduction theorem, $\mathbf{A} \vdash \mathbf{B}$ if $\mathbf{A} \to \mathbf{B}$, and $\mathbf{A}^- \supset \mathbf{B}^-$ is $(\mathbf{A} \to \mathbf{B})^-$. This applies if $\mathfrak{A}$ consists of negative formulae only.

At first sight the second method looks unpromising because of 3.302: in our paradigm, where $\mathfrak{A}$ consists of purely universal formulae, $\mathbf{A} \to \mathbf{B}$ is not in $\mathfrak{A}$ even if $\mathbf{A}$ and $\mathbf{B}$ both are. But it turns out to be *the* method of the subject.

3.32. *Cut Elimination in Predicate Logic.* For a nice list of cut free rules see [7]. The rules have two striking properties; after the corollaries we shall examine what *weaker* properties would be equally useful. (As stated in 3.242, *intuitionistic* logic is studied.)

3.321. A cut free proof of $\mathbf{A}$ contains only *subformulae* of $\mathbf{A}$, that is, logical components, or $\mathfrak{A}(\mathbf{t})$ if $\mathbf{A}$ is $\wedge x\mathfrak{A}(x)$, or $\vee x\mathfrak{A}(x)$ and $\mathbf{t}$ is an individual variable or constant.

3.322. There are *obvious* derived rules which are *not* valid as implications, for example,

3.3221. for negative (and somewhat more general [45]) $\mathbf{A}$, if $\vdash \mathbf{A} \to \vee x\mathbf{B}(x)$ then $\vdash \vee x[\mathbf{A} \to \mathbf{B}(x)]$, but not for all these $\mathbf{A}$, $\vdash[\mathbf{A} \to \vee x\mathbf{B}(x)] \to \vee x[\mathbf{A} \to \mathbf{B}(x)]$; cf. 2.34 concerning the axiom of choice.

3.3222. For quantifier free $\mathbf{P}(x)$, if $\vdash\neg\neg \vee y\mathbf{P}(y)$ or equivalently $\vdash\neg \wedge y\neg\mathbf{P}(y)$ then there is a finite sequence of terms $\mathbf{t}_1, \ldots, \mathbf{t}_n$ such that $\vdash\neg[\neg\mathbf{P}(\mathbf{t}_1) \wedge \ldots \wedge \neg\mathbf{P}(\mathbf{t}_n)]$ or, equivalently,

$$\vdash \neg\neg[\mathbf{P}(\mathbf{t}_1) \vee \cdots \vee \mathbf{P}(\mathbf{t}_n)].$$

Note the following connection with 3.3221. Take a finite set $\mathbf{A}_1$ of axioms for arithmetic which satisfies the conditions of [45] and implies $\mathbf{P}(\mathbf{t}) \vee \neg\mathbf{P}(\mathbf{t})$ for each term $\mathbf{t}$ and quantifier free $\mathbf{P}$. Then $\mathbf{A}_1 \to \neg\neg \vee y\mathbf{P}(y)$ if $[\mathbf{A}_1 \wedge \neg \vee y\mathbf{P}(y)] \to \vee y\mathbf{P}(y)$; take the premise for $\mathbf{A}$. Now cf. 2.322.

3.323. *First Main Result.* The *cut* rule is: from (i) $\mathbf{A} \rightarrow (\mathbf{B} \vee \mathbf{C})$, (ii) $(\mathbf{B} \wedge \mathbf{D}) \rightarrow \mathbf{E}$ infer (iii) $(\mathbf{A} \wedge \mathbf{D}) \rightarrow \mathbf{E} \vee \mathbf{C}$. There is a primitive recursive $\epsilon(p, q)$ such that, if p proves (i), q proves (ii) (without cut) then $\epsilon(p, q)$ proves (iii) without cut. This fact is also proved in primitive recursive arithmetic. The length of the proof $\epsilon(p, q)$ grows exponentially with the number of iterated implications and quantifiers in **B**. [31]

Corollary. Every theorem of ordinary intuitionistic predicate logic can be proved without cut.—N.B. *No assumption of completeness of the rules is made.*

3.3231. *Some Interpretations.* F predicate logic, F_0 intuitionistic propositional logic, F_1 the fragment of arithmetic needed for 3.323, $\mathfrak{A}$ purely universal, purely existential formulae or their double negations. By 3.3222 μ is an interpretation, if $\mu(n, \mathbf{A}) = \mathbf{A}$ if **A** does not contain existential quantifiers, if **A** is $\vee x\mathbf{P}(x)$, $\mu(n, \mathbf{A})$ is $\mathbf{P}(\mathfrak{t}_n)$ (in some enumeration of terms), if **A** is $\neg\neg \vee x\mathbf{P}(x)$, $\mu(n, \mathbf{A}) = \neg\neg[\mathbf{P}(\mathfrak{t}_1{}^n) \vee \cdots \vee \mathbf{P}(\mathfrak{t}_{k_n}{}^n)]$. This interpretation is *discriminating* (for $\mathfrak{A}$). It can be extended to all prenex formulae and their negations (no counterexample interpretation).

It does not seem to be known whether the whole of (intuitionistic) predicate logic has a discriminating interpretation in its quantifier free part.

If one regards classical logic F as the negative part of intuitionistic logic, one has trivially a discriminating interpretation (identity). But there is no *discriminating* interpretation of the latter in its negative part; this expresses in precise terms the impression that classical logic is essentially simpler.

3.3232. *Applications to Particular Axiomatic Theories.* Suppose, the axioms considered are quantifier free, that is, purely universal, and, on the intended interpretation, the function terms denote combinatorial rules. This holds for fragments of arithmetic in which the axioms are suitable recursion equations, and induction, applied to quantifier-free formulae, is a rule of inference; also for fragments of geometry, theory of algebraically closed fields applied to the field of algebraic numbers, and so on. Then 3.3231 shows essentially that the use of quantification theory in proofs of quantifier free formulae can be eliminated,[36] since 3.302 allows one to add universal axioms.

[36] The use of rules of inference: if quantification theory is not needed in the proof of the premise, apply the rule to the quantifier free theory, and start again.

This is a true reduction: for the intended meaning of the logical apparatus assumes *either* (in the classical case) the existence of infinite collections *or* the abstract theory of constructions in 2.2. In the cases above prime formulae are decidable, and so the propositional operations of 2.2 reduce to truth functions.

Whatever the verdict on Hilbert's program this result shows that *the mechanistic conception has a far wider application than appears on the surface.*

The results above extend naturally to all prenex formulae and their negations. Thus, for example, the formal independence of $\wedge x \vee y \mathbf{A}(x, y)$ can be shown without giving an interpretation of the axioms where it is refutable, but by showing that $\wedge x \mathbf{A}[x, \mathbf{t}(x)]$ is not satisfied by any of the *particular* function terms used in the interpretation.

3.3234. *Reflexiveness* (cf. 3.306). Here 3.321 is exploited. Suppose $F_0 \subset F_1$, and, for any given **A** there is an enumeration (restricted truth definition of 1.821) in F_1 of the subformulae of **A**. Further, relative to this enumeration the validity of the (cut free) rules of F_0 can be proved in F_1. Then F_1 is reflexive over F_0. (For careful statements in the classical case, cf. [53]; but the work carries over to the intuitionistic case except that not only the depth of quantifiers but also of iterated implications must be counted.)

What makes the observation useful is that for none of the usual F_1 is there a truth definition for all formulae, but often for subformulae of any given formula (subformula in the sense of 3.321);[37] for example, F_1 is first order arithmetic, F_0 predicate logic restricted to the notation of F_1. Following the convention of 3.3, boldface letters denote expressions of F_0, and when enclosed in corners terms of F_1 representing their numbers.

N.B. While the argument for 3.323 is purely equational (primitive recursive arithmetic), the proof of reflexiveness applies induction to $\mathrm{Prov}_0(m, \ulcorner \mathbf{A}[\mathbf{S}^n\mathbf{0}] \urcorner) \supset A(n)$ where $A(n)$ is the restricted truth definition, that is, in general to a quantified formula. Conversely, assume the schema 3.306 and proofs in F_0 of $\mathbf{P}(\mathbf{0})$, $\mathbf{P}(n) \rightarrow \mathbf{P}(n + \mathbf{1})$ with free variable n. Then, by elementary arithmetic: $\mathrm{Prov}_0[\sigma(n), \ulcorner \mathbf{P}(\mathbf{S}^n\mathbf{0}) \urcorner]$, and (by 3.306): $\mathbf{P}(n)$; that is, modulo this bit of arithmetic, *the schema* 3.306 *is equivalent to full induction.*

[37] Completely parallel work applies to quantifier free systems; only here truth definitions for formulae reduce to valuation functions for terms.

This type of equivalence will be generalized in 3.3321. Note that for Hilbert's original program, reflexiveness was of no interest except when **A** represents a combinatorial statement, when it reduces to consistency. But even for so-called ω-consistency it is useful, that is, with free variable **A** over formulae with the formal free variable $x:W(\mathbf{A})$

$$(n)\{\mathrm{Prov}_0[n, \ulcorner \rightarrow \wedge x\mathbf{A}(x)\urcorner] \supset \sim(m)(Ep)\ \mathrm{Prov}_0[p, \ulcorner \mathbf{A}(\mathbf{S}^m\mathbf{0})\urcorner]\}$$

For each fixed **A**, this follows from 3.306, that is, can be proved in F_1. To get ω-consistency itself, one only needs reflection *over* F_1 applied to the formula $W(\mathbf{A})$. Intuitionistically, ω-consistency is equivalent to (the negation of a prenex formula)

$$\sim(E\mathbf{A})(En)(m)(Ep)\{\mathrm{Prov}_0[n, \ulcorner \rightarrow \wedge x\mathbf{A}(x)\urcorner] \,\&\, \mathrm{Prov}_0[p, \ulcorner \mathbf{A}(\mathbf{S}^m\mathbf{0})\urcorner]\}$$

and thus can be interpreted in quantifier-free form provided 3.3231 is suitably extended; cf. 3.33.

3.33. *Cut elimination in arithmetic H* (2.412) is not possible in the following sense: there is no recursively enumerable sequence of *derived* rules with the following properties: (a) (for the canonical representation of the rules) it can be proved in H that the rules satisfy 3.321; (b) proof figures are finite so that induction can be applied as in 3.3234. For then we should have in H, for the derived proof relation $\mathbf{Prov}_D$: $\mathbf{Prov}_D(n, \ulcorner \mathbf{A}\urcorner) \supset \mathbf{A}$, contradicting 3.232. In fact, by 3.2321, (a) can be weakened to (a'): purely universal statements can be derived only from purely universal statements; and, of course, other variants.

Now, without (a), the cut free reformulation would lose much of its purpose. The clue for a proper weakening of (b) is this: to use *infinite proof figures*, already mentioned in footnote 28. The infinitary rule to be used is the so-called ω-rule: from $\mathbf{A}(\mathbf{0})$, $\mathbf{A}(\mathbf{1})$, . . . infer $\wedge x\mathbf{A}(x)$.

What restrictions should be put on the proof figures (if the resulting system is to be equivalent to ordinary arithmetic)? First, the obvious way of ensuring inclusion of the new system in the old is that one should be able to "talk" about the new system in the old. Precisely, each proof figure is a partial ordering of formulae with premises preceding conclusions: it should be definable in the old system, and it should be provable that (the formula at) any node is related to its immediate predecessors according to the rules. Finally, instead of (b), what is required is this: *transfinite* induction

be provable for the partial ordering, applied to

$$\mathbf{Prov}_I\,(m, \ulcorner \mathbf{A}(\mathbf{S}^n\mathbf{0})\urcorner) \supset \mathbf{A}(n)$$

where $\mathbf{Prov}_I\,(m, \mathbf{A})$ means: the formula $\mathbf{A}$ appears at the node m; and $\mathbf{A}(n)$ is the restricted truth definition. Second, there is the combinatorial problem of converting each proof of $\mathbf{A}$ in the ordinary system into an infinite proof in the new one. To be useful the infinite proof figure must have the properties just mentioned. ["Combinatorial" because quantifier free, that is, for a suitable function constant π: if $\mathbf{Prov}_H\,(n, \mathbf{A})$, $\pi(n)$ is (the number of) a description of an infinite proof tree for $\mathbf{A}$.]

3.331. *Basic Lemma.* Suppose an ordering $<$ is defined in (formal) arithmetic and there are terms $\mathbf{z}_0$, $\mathbf{z}_1$, $\mathbf{w}$, $\mathbf{s}(x, y)$, $\mathbf{exp}\,(x)$, which can be proved to satisfy the usual defining conditions for 0, 1, ω, ordinal addition, exponentiation to the base 2, and their inverses. (For precise statement, cf. [14].) Then, for $\mathbf{exp}\,(x)$ in the field of the ordering, transfinite induction (t.i.), in the sense of 2.623, up to $\mathbf{exp}\,(x)$ applied to a formula $\mathbf{Q}$ can be formally derived from t.i. applied to $\mathbf{Q}'$ up to x. $\mathbf{Q}'$ is defined from $\mathbf{Q}$ by means of a numerical quantifier. Further, there is an ordering (the first one thinks of) of ordinal ϵ_0 which satisfies the conditions above.

The proof in [14] is stated for classical arithmetic, but uses only intuitionistic rules. The particular conditions imposed on the terms $\mathbf{s}(x, y)$, $\mathbf{exp}\,(x)$ have this consequence. Suppose for another ordering $<'$, terms $\mathbf{z}_0'$ etc. can be proved to satisfy the same conditions, and the orderings restricted to predecessors of x, x' resp. have ordinal $\alpha(<\epsilon_0)$. Then the isomorphism between them can be proved formally.[38]

3.332. *Second Main Result* [14]. *Suppose there are cut free proof figures for* $\mathbf{A} \rightarrow (\mathbf{B} \vee \mathbf{C})$, $(\mathbf{B} \wedge \mathbf{D}) \rightarrow \mathbf{E}$ *which can be proved in H to satisfy the conditions of* 3.33. *Then one gets, primitive recursively in the description of these proof figures, a definition in H of a cut free proof figure of* $(\mathbf{A} \wedge \mathbf{D}) \rightarrow (\mathbf{E} \vee \mathbf{C})$. *The latter also satisfies the*

[38] To state this one needs function symbols, for example, H_1 of 2.412, which is a conservative extension of H. The result above is also of interest without the requirement of provability; for example, if the functions mentioned are recursive, the ordering is necessarily recursive, and so is the isomorphism. This intrinsic characterization of natural ϵ_0 orderings has been extended [23] to ordinals beyond ϵ_0 for suitable additional ordinal functions. It would be interesting to extend it to all recursive ordinals.

conditions of 3.33. *If the ordinals of the given proof figures* $<\alpha$, *the ordinal of the new one grows exponentially with the number of quantifiers and iterated implications in* **B**.

Corollary. Any proof of H can be reduced to a cut free proof of ordinal $<\epsilon_0$, that is, in the precise sense of 3.331. For, each use of induction in H can be replaced by a single application of the ω-rule; since there are only finitely many in any proof of H, we have an immediate reduction to a proof figure of ordinal $<\omega^2$ (with cuts).

Actually, the result states more than has so far appeared in the literature or (as far as I know) been done in detail. The work of [14] is formulated for the classical system, and somewhat informally only. There are indications for a more or less convenient *description* of infinite proof figures in [57]. The result stated above is what is needed for interesting applications (for example, reflexiveness of ϵ_0-induction over H).

3.3321. The consistency of H is proved by *applying transfinite induction on* ϵ_0 *in the following form* (*to a purely combinatorial statement*): for a certain primitive recursive τ, $\tau(n) \prec n \vee n = z_0$, $A[\tau(n)] \Rightarrow A(n)$, $A(z_0)$ are proved by combinatorial methods and then $A(n)$ is inferred. (To do this one arranges that the proof figures are *described* primitive recursively; this is not troublesome.)

3.3322. More generally, H together with ϵ_0-induction is *reflexive* over H. For, first, by quantifier-free ϵ_0-induction, Prov $(n, \mathbf{A}) \Rightarrow$ Prov$_{CF}$ $[\pi(n), \mathbf{A}]$ (for the cut free system), and then Prov$_{CF}$ $(m, \mathbf{A}) \supset A$. As in 3.3234, the reflexion schema is equivalent to ϵ_0-induction applied to arithmetic formulae.

Why can the use of transfinite induction be even contemplated as a *reduction* of ordinary induction? Ordinary induction is applied to *quantified* formulae which have, prima facie, no combinatorial meaning at all. The application of transfinite induction is at least *stated* in combinatorial terms.

3.34. *Refinements and Extensions.* The most significant extension to date, namely application to ramified analysis, will be treated below.

3.341. *Addition of Function Variables.* It seems plausible that cut elimination can be extended to H_1 or H^1 (2.51) with the full axiom of choice. The case of *free* function variables is implicit in [14], although the results are *stated* for function constants. For, all that is assumed in [14] are axioms of the form $\alpha(\mathbf{S}^n\mathbf{0}) = \mathbf{S}^m\mathbf{0} \vee$

$\alpha(\mathbf{S}^n\mathbf{0}) \neq \mathbf{S}^m\mathbf{0}$. The appropriate additional rule is: infer $\mathbf{A}(\alpha)$ from $\alpha(\mathbf{S}^n\mathbf{0}) = \mathbf{0} \rightarrow \mathbf{A}(\alpha)$, $\alpha(\mathbf{S}^n\mathbf{0}) = \mathbf{S0} \rightarrow \mathbf{A}(\alpha)$, and so on.

In a cut free proof of $\forall x\mathbf{A}(\alpha, x)$, the only way of introducing the existential symbol is then: $[\alpha(\mathbf{0}) = \mathbf{S}^{n_0}\mathbf{0} \wedge \cdots \wedge \alpha(\mathbf{S}^m\mathbf{0}) = \mathbf{S}^{n_m}\mathbf{0}] \rightarrow \forall x\mathbf{A}(\alpha, x)$ from $[\ldots] \rightarrow \mathbf{A}(\alpha, S^p\mathbf{0})$ for some fixed $p, n_0 \ldots, n_{n_m}$. A consequence, already referred to on page 143: if $\mathbf{A}(\alpha, x)$ is monotone and $\forall x\mathbf{A}(\alpha, x)$ is *proved* in H^1 (with $\mathbf{A}$ in place of R in 2.6231) then bar induction is also provable in H^1 (2.6231).

3.3411. The inclusion of free function variables permits the extension of 3.3231 (no counterexample interpretation) by applying transfinite induction to formulae containing free function variables. Let $\mathbf{A}$ be the negation of a prenex formula:

$$\neg\wedge x_1 \forall y_1 \ldots \wedge x_n \forall y_n \mathbf{R}(x_1, \ldots, x_n, y_1, \ldots y_n).$$

Then

$$\mathbf{A} \rightarrow \neg\wedge x_1 \ldots x_n \mathbf{R}[x_1 \ldots x_n, \alpha_1(x_1), \ldots, \alpha_n(x_1 \ldots x_n)]$$

which we write $\neg\wedge x\mathbf{S}(\alpha, x)$, where $\mathbf{S}$ is quantifier-free. But the *derived* rule 3.3222 carries over to arithmetic:[39] so $\forall x \neg\mathbf{S}(\alpha, x)$ is also provable. Apply 3.341.

3.342. *The Significance of* ϵ_0. As shown in [14], if instead of starting with the schema of ordinary (ω) induction, one starts with induction on an arbitrary ordering α, one gets to the next ϵ-number past α (for orderings satisfying the conditions cited in 3.331).

3.3421. For negative results much weaker conditions are needed. Evidently, by 3.3321, for orderings of ordinal ϵ_0 satisfying the conditions of 3.331, transfinite induction cannot be proved in H (even for certain primitive recursive predicates). But much more is true, by use of 3.341.

Suppose $x \prec y$ is an *arithmetic* ordering (or even one defined, in prenex form, with existential function quantifiers only), and transfinite induction can be proved for it for all formulae containing free function variables. Then its ordinal is $< \epsilon_0$.

Consider the statement which expresses that α is an order pre-

[39] So does 3.3221: this wholesale carry over of derived rules from predicate logic is one of the special virtues of cut free infinite proofs. An important exception is this: if $\mathbf{A} \vee \mathbf{B}$ is provable in predicate logic either $\mathbf{A}$ or $\mathbf{B}$ is provable. In arithmetic, only if $\mathbf{A}$ and $\mathbf{B}$ are *closed*.

serving mapping of the natural ϵ_0 ordering into a segment of $x \prec y$. It has the form: $\wedge u\mathfrak{A}(\alpha, u)$ where $\mathfrak{A}$ is arithmetic. By 3.3321

$$\mathbf{Prov}\,[n, \ulcorner \neg \vee \alpha \wedge u \mathfrak{A}(\alpha, u) \urcorner] \rightarrow \neg \vee \alpha \wedge u \mathfrak{A}(\alpha, u)$$

is proved by induction on ϵ_0 applied to a particular formula $\mathfrak{A}(\alpha)$. But from $\wedge u\mathfrak{A}(\alpha, u)$ together with (provable) transfinite induction on $\prec$ applied to a suitable $Q'(\alpha)$, we have induction on ϵ_0, and hence $\{\wedge u\mathfrak{A}(\alpha, u) \wedge \mathbf{Prov}\,[n, \ulcorner \neg \vee \alpha \wedge u \mathfrak{A}(\alpha, u) \urcorner]\} \rightarrow \neg \vee \alpha \wedge u \mathfrak{A}(\alpha, u)$, that is, $\mathbf{Prov}\,[n, \ulcorner \neg \vee \alpha \wedge u \mathfrak{A}(\alpha, u) \urcorner] \rightarrow \neg \vee \alpha \wedge u \mathfrak{A}(\alpha, u)$. But then, by 3.233, we have a proof of $\neg \wedge u \mathfrak{A}(\alpha, u)$. (There is a corresponding argument for the cut free system, only now ϵ_0-induction is needed for a formal proof of closure under cut.)

It is essential to demand that t.i. on $\alpha \prec y$ be proved for formulae containing *free* function variables since there are orderings which are not even well orderings, yet t.i. can be proved for all *arithmetic* predicates Q.

3.343. Cut elimination and consistency proofs. 3.3321 makes it clear that the foundational significance of consistency proofs by transfinite induction depends on the justification of the latter, and 3.331 suggests that this justification should be in terms of the way the ordering is built up. The following quite trivial remarks summarize the *formal* situation. To fix ideas, consider formal systems of analysis.

Consistency can be proved by t.i. on a primitive recursive well ordering of ordinal ω, simply because it is an universal statement. [$\wedge u A(u)$ can be proved by t.i. on the following ordering: $x \prec y$ if either $x < y \wedge (\wedge u < x) A(u)$ or $y < x \wedge \neg(\wedge u < y) A(u)$.] If we take the union of all primitive recursive orderings which can be proved to be ω orderings in say analysis, the consistency can be proved by induction on this ω^2 ordering. For every finite subsystem can be proved to be consistent, and so, by induction on one of the orderings in the sequence.

The situation is quite different if one examines cut elimination itself. Suppose we demand 3.33 (a) and the validity of the rules to be proved in elementary analysis, and instead of (b) merely that the proof figures be recursive and well founded. Then the bar theorem (2.623) even for *recursive* R does not have a cut free proof. Since 3.33 (a) was essential to the more interesting applications of cut elimination, a generalization of the notion of cut elimination for analysis has quite a different character [72].

3.35. *Other Approaches.* The use of finite types (2.321) has been described at length in the previous section. It can be taken as a justification of transfinite induction up to ϵ_0 or else be regarded computationally and reduced to t.i. (2.35).

3.351. Another method, particularly natural for the classical system, is Hilbert's ϵ-substitution method, where quantifiers are replaced by so-called ϵ-symbols. A very elegant and mathematically satisfying treatment is given in [68]: it presents a reformulation parallel to 3.24. Hilbert's substitution problem requires the replacement of ϵ-terms by numbers to satisfy a finite set of purely numerical conditions; the possibility for doing this is ensured by the intended meaning. Hilbert introduced a particular sequence of approximations: to show that this converges *is* a problem, but, in the case of arithmetic, easily settled by nonconstructive continuity considerations. In [68], the problem is generalized to a class of functional equations and one seeks to determine *how the solutions depend on the parameters of these equations.* The *natural* solution of this problem is constructive and makes the appearance of ϵ-numbers very clear.

Actually, in [17] a sequence of approximations was defined for the case of analysis. The *proof* of convergence given was defective. It does not seem to be known (even nonconstructively) if the method converges at all.

3.352. By extension of 3.351, there is a direct treatment of the no-counterexample interpretation for classical arithmetic [69]. In fact, to date it is the only detailed treatment for the assertions of 3.3411. It applies to the intuitionistic case since the negation of a prenex formula is provable in H if and only if it is provable classically (by 3.31 and 2.33).

3.4. *Finitist Proofs.* Perhaps the most satisfactory kind of theory of finitist proof, or any other kind of reasoning, would take this form: properties of these proofs are formulated in terms of a basic conceptual framework, such as (some extension of) the theory 2.2, properties which are sufficient to characterize the notion or, at least, to decide specific questions about them. This would correspond to a *definition* of the class of hereditarily finite or hereditarily countable sets in abstract set theory (cf. also 1.32 or 1.835). Such a *theory about* finitist proofs would be independent of cumbersome notation (cf. 3.4144). The present section, which gives a *description of* finitist proofs, is at the other end of the scale

and uses ad hoc devices. One reason is that the known abstract theories are not rich enough to express these devices. But a possible independent interest of the present section is that its formal derivations reflect more closely the complexity of finitist proofs, in contrast to a theory about them.

3.41. About finitist proofs. Explanations of the intended meaning are given on page 119 or 3.111 in terms of the distinction between concrete and abstract operations; less technically, one speaks of proofs that one can *see* or *visualize*. The latter is more vivid, but misleading unless one remembers that our primary subject is a *theoretical* notion for the analysis of actual visualizing, not that experience itself.[40]

Warning. The name "finitist" was introduced by Hilbert because he thought that finiteness was the essential feature of the elementary kind of evidence (cf. 3.111) he had in mind. While it is clear that the *property* of proofs be stressed is significant, it should not be assumed that his *analysis* was correct. The analysis is not plausible: he wanted to go beyond old-fashioned formalism; for this it was essential to have *general* assertions like the commutative law, *not* confined to a finite domain, which are capable of elementary evidence.

The formalisms below are quantifier-free systems with variables over natural numbers $(n, m, p, q, \ldots)$, constants for number theoretic functions $(a, b, c, \ldots)$ introduced by defining equations, and propositional combinations of equations. Also there are some essential *auxiliary* expressions (cf. 3.2341). Primitive recursive arithmetic (PR) is a simple example.

3.411. *The meaning of:* $a(n) = b(n)$. First of all, if such a statement is finitistically proved, it is *true* on the (corresponding) interpretation of Section 1; also it is constructively proved in the sense of Section 2. But these consequences are here purely incidental: the restriction to finitist reasoning is quite unnecessary for them,

[40] Perhaps, related to what can actually be visualized (by a particular person at a particular time), as the theoretical (mathematical) notion of: constructible by ruler and compass is related to: practically constructible by means of some particular ruler and compass, etc. Note at once: (a) The notion below is crude because it does not give a realistic measure of complexity for the (visualized) structures considered; it uses ordinals $< \epsilon_0$: how much bigger is ω^{ω^ω} than ω^ω? (b) The work is about proofs concerning *discrete* structures, typically: arithmetic; whether evidence, that is, seeing, in geometric reasoning requires an essentially different analysis is left open.

and it is precisely this restriction which is our main subject. Instead, $a(n) = b(n)$ expresses that one can (theoretically) visualize the whole computation procedures for $a(0)$, $a(S0)$, . . . and $b(0)$, $b(S0)$, . . . and convert them into one another by restructuring. A justification of formal rules consists in convincing oneself that a formal proof of $a(n) = b(n)$ enables one to do just this. Evidently, the use of combinators would be much more suitable than our use of equational systems; certainly, even the rule of substitution requires analysis for the present interpretation.

3.412. *Definition Principles.* Recall the section on Church's thesis. The analysis of computations makes it quite clear that all finitist definitions of number theoretic functions can be reduced to the form of *recursion equations*, that is, this: A propositional combination $E(f, n)$ of equations built up from the function letter f, constants, and sets $\mathfrak{n}$ of variables $(n_1, \ldots n_n)$ such that, for each numeral S^n0, the value of $f(S^n0)$ is *finitely determined.*[41] To be finitist this fact must be finitistically recognized. Suppose a collection $\mathfrak{E}$ of such accepted equations together with the structure of all their computations has been visualized. This assumption is formally expressed by introducing a valuation function $\vee_{\mathfrak{E}}$ for terms built up from the symbols of $\mathfrak{E}$, with the obvious equations for $\vee_{\mathfrak{E}}$ (footnote 37): if a_0 is a constant in $\mathfrak{E}$ with Gödel number $\bar{a}_0$, and the term $a_0(S^m0)$ has number $s(\bar{a}_0, m)$, the term $a_0(b)$ has number $r(\bar{a}_0, \bar{b})$ then $\vee_{\mathfrak{E}}[r(\bar{a}_0, \bar{b})] = \vee_{\mathfrak{E}}\{s[\bar{a}_0, \vee_{\mathfrak{E}}(\bar{b})]\}$, and similarly for each equation in $\mathfrak{E}$.

3.413. *Learning to Visualize.* The basic equipment assumed is this. If one can visualize a structure, then also a sequence of ω copies; more generally, if a is a function already introduced, and s a structure, the sequence s, $a(s)$, $a[a(s)]$, . . . can be visualized. Evidently, it is understood that the latter is *given* as this sequence, and not merely in extension, since otherwise the previous exercise of visualizing s would not help.

3.414. *Iteration beyond* ω. For stating the facts we use notations for ordinals $< \epsilon_0$ built up from 1, ω, $+$, $\cdot$ and exponentiation; Greek letters ξ, η, . . . are variables for them, and bold face for particular figures. The choice will explain itself; concerning the properties of this particular notation needed below, cf. 3.31. By

[41] For some substitution $\mathfrak{n}_0$ of numerals and m, $E(f, \mathfrak{n}_0) \vdash f(S^n0) = S^m0$. This latter relation is decidable for fixed n, m, $\mathfrak{n}_0$. The mere fact of finite determination is not characteristic for finitist rules but only for recursiveness.

above (3.413), if an iteration (of some operation) up to ξ has been visualized, then it can also be visualized to $\xi \cdot \omega$. So much is clear. Equally it is clear that it may so happen that iteration to each $\xi < \xi_0$ can be visualized (with more and more effort!) but not ξ_0 itself. What one would like to say is this: *if one sees how to visualize iteration up to each* $\xi < \xi_0$, then ξ_0 *itself can be visualized.* The problem is to put this into formal terms. Or, more vividly, can we find a formal expression $L(\xi)$ such that a formal proof of $L(\xi)$ by suitable rules enables us to visualize iteration up to ξ?

3.4141. Throughout "iteration up to ξ" means iteration from $\mathfrak{S}$ to $\vee_{\mathfrak{S}}$ and primitive recursive definitions in $\vee_{\mathfrak{S}}$ (3.412).

3.4142. As always in case of doubt, one has to treat positive and negative questions differently, namely that each $\xi < \epsilon_0$ can be theoretically visualized, and none beyond. *The latter is to be done by finding suitable closure (inaccessibility) properties of* ϵ_0.

3.4143. We shall introduce a symbol (pseudoproperty) O applied only to particular figures ξ, and not used for building up composite formulae. The intended meaning of $O(\xi)$ is: ξ can be visualized (as in 3.4141). Since the totality of all (concrete) structures which can be visualized, cannot be visualized, we do not use variable ξ, though bounded variables would be possible. More important, negations of (hence implications with the premise) $O(\xi)$, even for constant ξ, would refer again to a totality that cannot be visualized, and are therefore excluded altogether.

3.41431. (Cf. footnote 40.) For a more realistic theory of actual practice the step from ξ to $\xi \cdot \omega$ may have to be restricted for larger ξ. A little experimentation shows that even for relatively small ξ it is not natural to visualize the configurations involved. Put differently: even when we could prove an assertion finitistically, in practice we do not.

3.4144. It is possible that a more elegant theory would use logical connectives and quantifiers, an interpretation in terms of the basic concepts here used, and rules valid for it. It is not excluded that the simple theory 2.21 has this property, that is, is valid if one interprets $\pi[a, x \cdot v(x)]$ as: a is a finitist proof of $v(x) = 0$. But the matter is delicate, and so we avoid it altogether in the

3.42. *Two Formal Systems.* The general idea is this: start with primitive recursive arithmetic, which is supposed to have been thought through finitistically. Therefore 3.412 can be applied, and this step is to be iterated. As we go on we carry along $O(\xi)$ to

indicate that ξ has been visualized and so the step 3.412 is iterated ξ times. Symbols P_ξ, resp. $P_\xi{}^+$ are introduced provided $O(\xi)$ has been established, and $P_\xi(n, m)$ is to express that n represents a finitist proof of the formula m. (Formula is one that does not contain O, pseudoformulae may contain O.) Some reasonable arithmetization of the syntax is assumed.

Both systems contain the usual schemes of primitive recursive arithmetic, and a supply of function constants. (It will turn out that the P_ξ are definable. But in the first place they are intended to be used in axioms *about* proofs by use of ξ iterations of 3.412.)

P_0 is an explicit proof predicate for PR.

$O(\omega)$ is an axiom. And we infer $O(\eta \cdot \omega)$ from $O(\eta)$, and for $\xi < \eta$, $O(\xi)$ from $O(\eta)$.

3.421. Suppose $O(\xi)$. Then $P_\xi(p, q)$ is equivalent to the disjunction of conjunctions:

3.4211. $p = \langle \eta, p_1 \rangle$, $\eta < \xi$, $P_\eta(p_1, q)$ (accumulation).

3.4212. $p = \langle p_1, p_2 \rangle$, $P_\xi(p_1, q_1)$, $P_\xi(p_2, q_2)$, and q is a consequence of q_1 and q_2 in PR (validity of rules of PR).

3.4213. $p = \langle \eta, p_1 \rangle$, $\eta < \xi$, q is a recursion equation for the valuation function of terms in P_η, $\eta < \xi$ (cf. 3.412).

3.4214. (for handling O), $p = \langle \eta, p_1 \rangle$, $\eta < \xi$, $q = O(\zeta)$ and $P_\eta[p_1, \ulcorner \xi < \zeta \rightarrow P_\eta\{a(\xi), \ulcorner O(\xi) \urcorner\} \urcorner]$ (with formal variable ξ; note that O occurs only with fixed arguments), and, of course, $p = \langle p_1, \zeta \rangle$, $q = O(\zeta \cdot \omega)$, $P_\xi(p_1, \ulcorner O(\zeta) \urcorner)$.

3.4214 is intended to express that one has finitistically recognized that each $\xi < \zeta$ can be visualized.

Finally: if $P_\xi(n, \ulcorner \mathbf{A} \urcorner)$, then $\mathbf{A}$.

3.422. It is clear that for just the ordinals $\eta < \omega^{\omega \cdot \xi}$, $O(\eta)$ is provable in the system P_ξ.

3.423. In the second system condition 3.4214 is *relaxed* so that in the first place it is not even a formal system.

$P_\xi{}^+(p, q) : q = O(\zeta)$ and for each $\eta < \zeta$, $\vdash_\tau O(\eta)$ is true for some $\tau < \xi$.

(This is consistent with 3.414, because it is required that $\tau < \xi$.)

3.43. *Main Result.* $O(\xi)$ *both in* 3.421 *and* 3.423 *just for* $\xi < \epsilon_0$. *By* 3.422, *the same formulae are theorems of the two systems.*

3.44. *Simplifications for mathematical practice* (cf. 1.2 or footnote 14). While the detailed analysis of the process of finitist reasoning requires something like the hierarchy of proof predicates P_ξ, in practice one wishes to suppress them. One way is this: by 3.43

and the basic results of 3.32, the system H (of intuitionistic) or, equivalently, Z (of classical arithmetic) provides such a simplification.

3.441. *For each theorem* $\wedge x \vee y \mathbf{A}(x, y)$ *(with* $A \in$ PR*) of* H *or* Z, *there is a function constant* a *in* 3.421 *such that* $A[n, a(n)]$ *in* 3.421. *Conversely, for each constant* a *in* 3.421 *there is an explicit definition in* H *(for which the defining equations of* a *can be formally proved).*

3.4411. The result extends immediately to conjunctions and disjunctions of *prenex* theorems of H. These are the only formulae of H to which an *immediate* finitist sense is attached. For extensions to other formulae one needs interpretations, cf. 3.3411.

3.442. Perhaps a more attractive formulation is this. One adds to PR free function variables and a constructive *existential* numerical quantifier with the obvious rules. Further, the inference rule, for relation constants R:

If $(Ex) \sim \alpha(x + 1) R\alpha(x)$ infer transfinite induction on R applied to existential formulae.

This cannot be expressed in its full form:

$$\text{from} \quad (x)[(y\ R\ x)(Ez)A(y, z) \cdot \supset (Ez)A(x, z)] \qquad \text{infer } (Ez)A(x, z),$$

but instead, from

$$(Ez_1, z_2)\{A[a(x), z_1] \,\&\, [a(x)Rx] \,\&\, A[b(x, z_1), z_2] \,\&\, [b(x, z_1)Rx]\} \supset (Ez)A(x, z)$$

infer

$$(Ez)A(x, z).$$

However, the latter is sufficient for so-called *nested recursion* on R, and thus, by [68], one gets again to ϵ_0.

3.4421. One would like to interpret the free function variables as variables over *finitistic* functions (not free choice sequences). The mere truth of $(\alpha)(Ex)[\sim\alpha(x + 1) R\alpha(x)]$ would not be sufficient to justify the inference; for, if finitistic functions are assumed to be in a recursively enumerable class E of recursive functions, there are R_E such that $(\alpha \in E)(Ex)[\sim\alpha(x + 1) R_E\alpha(x)]$ but the inference is false (for suitable R_E). On the basis of 3.42, we know that a finitistic *proof* of the premise is in fact sufficient for the conclusion, but there seems to be no direct way of seeing this. Cf. in 3.534 the direct justification of the hyperarithmetic comprehension rule.

3.45. We defer the discussion of open problems on finitist proofs to the end of the section because most of them also arise with

3.5. *Predicative Proof.* Consider the variant of the ramified hierarchy of 1.5 for definitions of sets of natural numbers only. The first step is this: given the collection N (of natural numbers), the collection D of formal *definitions* of subsets of N by means of the usual logical operations is well defined, and hence the corresponding collection N_1 of subsets of N defined by D.[42] The general step is to apply this process to a collection N_ξ instead of N. The formal expression of this idea is exactly parallel to the introduction of a valuation function in 3.412: here, given a collection D, we introduce instead a *truth definition* T_D for sentences (from which a satisfaction definition is derived by use of a substitution function s as in 3.412).

Again the basic problem is: how often is this step to be iterated (cf. 1.52). There are some formal similarities to the iteration of 3.414. But the basic difference is this: in the former, iteration has to be *finitistically* justified, here *predicatively* or: on the assumption that *only the collection N is presupposed* (reducibility to N). Different formal axioms for (what corresponds to) $O(\xi)$ (above) may be expected to follow from, that is, express the operational significance of, these two requirements.

In 3.4 some attempt was made to say something about the intended meaning of finitistic justification (in terms of visualization of computations); in fact, if ϵ_0 is going to be the bound, it is to be expected that all the technical apparatus is in 3.4 already: For predicative justification, the technical apparatus has only recently been created; it is described very fully and attractively in the recent [29]. It will be sufficient to report the results briefly.

Here it should be remarked that there are also some formal differences between the two cases; for example, the step 3.414 is very *sensitive* to the *description* of $\mathfrak{E}$, as would be expected: it depends *how* we see the structure of computations, cf. [27]. The present basic step not, as seen from the notation free formulation in 1.5.

3.51. A *definition* of: *β-numbers* (abstract ordinals): 0-numbers are the powers of 2 ($\chi_\alpha{}^0 = 2^\alpha$); for $\beta \neq 0$, β-numbers α satisfy $\chi_\alpha{}^\gamma = \alpha$ for all $\gamma < \beta$. Γ is the least ordinal γ for which $\chi_0{}^\gamma = \gamma$.

Existence of such ordinals follows from known results. For positive results, as in 3.31, it is necessary to have orderings of ordinals $< \Gamma$ on which certain basic ordinal operations can be defined

[42] For the classical interpretation of the logical operations if the idea is that the *collection N* is given.

arithmetically and proved to satisfy their "recursive definitions," cf. [29].

In one respect the present theory is easier than 3.4: there are meaningful statements of predicative mathematics which are quite "close" to the impredicative idea of: R is well founded, for instance: if $\mathrm{TI}(R, P)$(TI for transfinite induction) stands for

$$\forall x\{[(\forall y\, R\, x)P(y)] \Rightarrow P(x)\} \Rightarrow \forall x P(x)$$

and P^ξ is a variable over sets of the hierarchy at level ξ, take $\forall P^\xi \mathrm{TI}(R, P^\xi)$. Now, this only expresses TI for a restricted class of predicates, and so, as a premise of an *implication*, it does not have the full force of well foundedness. But used in a *rule* of inference, it does: for, a *formal proof* of even $\forall P^1 \mathrm{TI}(R, P^1)$ can be used as a *schema* of proofs for $\mathrm{TI}(R, \mathfrak{P})$ for an arbitrary $\mathfrak{P}$ [in this sense: suppose the highest level occurring in $\mathfrak{P}$ is η, and the proof of $\forall P^1 \mathrm{TI}(R, P^1)$ assumes the existence of ξ levels; then the existence of $\eta \oplus \xi$ levels ensures $\mathrm{TI}(R, \mathfrak{P})$].

3.52. *Main Result. For a natural formal theory of the ramified hierarchy T_ξ, when T_ξ is used only if $\forall P^1 \mathrm{TI}(\xi, P^1)$ has been proved in T_η for some $\eta < \xi$: The length of the hierarchy is precisely* Γ [29].

3.521. The conclusion is shown to be insensitive to the exact formulation of the idea of well ordering. But more important

3.522. *The bound Γ is unaffected if the rules of proof in T_ξ are enlarged; namely: η applications of the nonconstructive ω-rule are permitted provided $\forall P^1 \mathrm{TI}(\eta, P^1)$ has been proved.*

3.523. Γ is also the precise bound if one uses instead of the ramified hierarchy of subsets of N the more elegant full hierarchy of 1.5, with the corresponding autonomy condition.

3.53. *Simplifications.* One of the most satisfying results of [29] is a beautifully elegant equivalent formulation IR (inductive recursion) of the theory $\bigcup_{\eta<\Gamma} T_\eta$; (IR) is an ordinary second order theory (in which the ramification is suppressed), (a) a subsystem of classical analysis, but (b): For each theorem of (IR) there is an $\eta < \Gamma$, such that the theorem holds in N_η. In other words, by (a) *all theorems of* IR *are valid for classical analysis,* but also by (b) *justified on the assumption of N only.* It can be verified that the natural formalization of the bulk of classical analysis only uses the axioms of IR.

It should be noted that the essential axiom of IR is the *rule* of inference corresponding to the *implication* of 2.623 (bar theorem).

3.531. We give here a formulation for *intuitionistic* logic which is proof theoretically equivalent to IR. (It is slightly smoother than IR itself because we can use the unrestricted axiom of choice instead of special definition principles; cf., for example, 2.421.) This has some independent interest (2.423); but note that, for the interpretation of Section 2, this theory does *not* have predicative character at all, because the intuitionistic logical operators are themselves defined self-reflexively.

Add to first order arithmetic H (2.412) predicate variables S, T (for species of natural numbers), and the usual formation rules. Extend the schemata of H to the new formalism. [$S(x)$ is not assumed decidable: in contrast to the case of function variables in H_1 or H^1.]

3.5311. The axiom of choice $(n)(ES)A(n, S) \supset (ES_1)(n)A_1(n, S_1)$ for arbitrary formulae A, where $A_1(n, S_1)$ expresses that S_1 is a choice set.

3.5312. A definition principle: *hyperarithmetic comprehension rule.* Suppose A and B contain no predicate quantifiers, but possibly several free predicate variables; also suppose for the formula $P(n)$ the following conditions are proved

$$(ES)A(n, S) \supset P(n), \qquad (ET)B(n, T) \supset \neg P(n), \text{ and}$$
$$\neg\neg[(ES)A(n, S) \vee (ET)B(n, T)].$$

Then we introduce the axiom, if U does not occur in $P(n)$,

$$\neg\neg(EU)(n)[U(n) \leftrightarrow P(n)].$$

3.5313. An inference principle: suppose the formula R is decidable $[R(n, m) \vee \neg R(n, m)]$ and for predicate *variable* U: $\mathrm{TI}(R, U)$ is proved. Then, for *arbitrary* formulae A, $\mathrm{TI}(R, A)$ may be inferred.

3.5314. In contrast to 3.4421, these axioms express some evident properties of the intuitionistic ramified hierarchy I_η of sets of natural numbers. (The argument for the classical case is parallel.) The assertion is that any theorem of the system holds when the predicate variables are interpreted to range over some suitable I_η *and beyond.* 3.5311 is clear since if the premise is proved for I_η, the conclusion for $I_{\eta+1}$. Suppose the premises of 3.5312 are *proved* for I_ξ, $\xi \geq \eta$; then (a) $\neg\neg[(ES)A(n, S) \leftrightarrow \neg(ET)B(n, T)]$, that is, for S, T ranging over I_ξ $(\xi \geq \eta)$; (b) $(ES_\eta)A(n, S) \supset (ES_\xi)A(n, S)$. (c) $\neg(ET_\xi)B(n, T) \supset \neg(ET_\eta)B(n, T)$. Hence $\neg\neg[(ES_\xi)A(n,S) \leftrightarrow (ES_\eta)A(n, S)]$ and, by the conditions on P, $\neg\neg[P_\xi(n) \leftrightarrow$

$P_\eta(n)$]. Take $P_\eta(n)$ for U. Finally, the inference principle is justified by using the *formal* proof of TI(R, U) as a *schema* (cf. 3.51) for proving TI(R, A). Note that these considerations would not apply to the corresponding *axioms* instead of rules (cf. 3.2341).

3.6. *Consequences of* 3.4 *for Hilbert's Original Finitist Program, of* 3.5 *for the Analogous Predicative or Nominalist*[43] *Program.*

3.61. On the positive side. The basic steps in the hierarchies 3.4–3.5 are clear, and "practical" proofs [in full arithmetic Z, resp. in (IR)] use only a few of the permitted rules. So, if they are formalized at all in these systems, they can be reduced to the first few steps of the corresponding hierarchy, and *therefore are unaffected by the problem of iteration.*

3.62. On the negative side. It is certainly true that (a) in crude finitism one contemplated only proofs that are *immediately* formalizable in 3.42, particularly 3.423, and (b) in crude predicativity, proofs that are obviously formalizable in 3.52, particularly 3.522. So, at least as originally conceived, Hilbert's program cannot be carried out.[44] (Specifically, in 3.42 one cannot prove the consistency of first order arithmetic H, in 3.52 one cannot prove the consistency of the bar theorem 2.63; further, more conclusively, in 3.423 one cannot prove the consistency of H; of course, all true arithmetic statements can be proved in 3.522, but *there are statements* TI(R, P) *which can be proved by means of the bar theorem but not in* 3.522.)

3.621. The only support for taking ϵ_0, resp. Γ as bounds is empirical: different formulations lead to the same ordinal; and for ϵ_0, Γ as *upper* bounds that the bounds are not extended by replacing some of the rules by their nonconstructive versions. This kind of support is subject to the criticism of 2.73. Formally speaking, the difficulty is this: for example, all axioms for P_ξ in 3.4 assert interrelations or the inference of a positive statement about P_ξ, in other words, just *closure* conditions on the class of finitist proofs.

[43] Again in a suitable theoretical sense, as in footnote 40. There is some marginal literature on a curious empirical nominalism, where mathematical assertions are to be interpreted in terms of material inscriptions and their (spatio-temporal) relations. But the work of 3.5 seems a more fruitful development of the ideas implicit in crude (nonconstructive) nominalism.

[44] Note that it was not clearly understood till [32] that the then *current* finitistic proofs could be easily formalized in Z, for example, the use of nested recursion or equations involving higher type variables.

This compares badly with the axiomatic theory of *validity* in 1.8 where, for example, 1.833 allows the *refutation* of a statement about validity by proving a statement not involving this notion. Non-technically, one can express it this way: what "more" do I know about an identity $a(n) = b(n)$ if I know that it is finitistically proved and not merely true? If one could express convincingly this extra information (even partially) in ordinary arithmetic terms, one would have the analogue to 1.833. All we offer is: formal provability in 3.42, and the completeness of these rules for finitistic provability is certainly not immediately convincing; cf. 1.832. It may be remarked that the completeness of the usual rules of logic with respect to logical validity was certainly practically certain for a long time, but a satisfactory theoretical foundation required that the known facts were pulled together and formulated in the basic completeness theorem (1.834). It is the analogue to this last step that is badly missing in 3.4. (For 3.5 the situation is better because predicative provability certainly implies *truth* in the hierarchy up to ω_1, cf. 1.533: this allows one to be sure that, for example, the theorem of Cantor-Bendixson cannot be proved predicatively [29], cf. 2.63.)

3.63. To avoid misunderstanding, two points should be noted.

3.631. Both ϵ_0 and Γ are "finally" reached as the limit of an ω sequence, for example, $\epsilon_0 = \lim \omega_n$, where $\omega_1 = \omega$, $\omega_{n+1} = \omega^{\omega_n}$. "Since ω sequences are permitted and each ω_n is permitted, why not ϵ_0?" One must not forget what is involved in the assertion that ω_n is permitted, namely: ω_{n-1} iterations of the basic principle of 3.4 were needed to introduce ω_n; so $\Sigma_{n=2}^{\infty} \omega_{n-1}$ steps are needed to introduce the *sequence* ω_n, that is, ϵ_0 steps. One has to know that ϵ_0 itself is permitted before it can be seen to be reached by the basic step of 3.4. Similarly for Γ: a typical impredicativity (relative to the basic steps considered).

3.632. Both in 3.4 and 3.5 much larger orderings than Γ can be defined, which are, impredicatively speaking, well orderings. (In fact, the same bound ω_1 for both by [65].) Briefly, these are: well-founded finitist, resp. predicative orderings. What was analyzed above were the notions of: *finitist,* resp. *predicative well orderings.*

3.6321. Recall 1.533: If, in the construction of the hierarchy, T_ξ is introduced provided for some $\eta < \xi$, there is a well-founded T_η-ordering of ordinal ξ (not necessarily: a T_η-well ordering) the catching point is ω_1: so ω_1 "corresponds" to Γ. It would be interesting

to know what ordinal corresponds to ϵ_0, if the step 3.412 is continued for suitable *natural* (recall [27]!) well orderings, perhaps in the sense of footnote 38.

4. IMPREDICATIVE (FULL) CLASSICAL ANALYSIS

A problem in formal proof theory (mainly for specialists). In the present section are collected a number of essentially equivalent formulations of classical analysis, which were obtained in the course of extending the interpretation 2.32 to full impredicative analysis.[45] They have some technical interest. But, more important, they provide basic problems on constructive functions of finite type which test, and thereby further, our understanding of these objects. The properties needed in Section 2 were too elementary for ambiguities, for example, those of 2.331, to matter.

Throughout this section lower-case letters denote numerical variables, capitals denote variables over sets of natural numbers, α, β denote sequences of numbers or type τ objects.

4.1. *Basic Systems for Classical Analysis* ([6] or [66]); all systems considered contain Peano's axioms.

4.11. The full *comprehension schema:*

$$\wedge S \vee T \wedge x[x \in T \leftrightarrow A(x, S)]$$

for *arbitrary* formulae A, not containing T, and classical logic.

4.12. *The Principle of Dependent Choices* (The strongest form of the axiom of choice expressible in the present notation):

$$\wedge \alpha \vee \beta A(\alpha, \beta) \rightarrow$$
$$\wedge \alpha \vee \beta_1 \wedge x\{\beta(<0,x>) = \alpha(x) \wedge A[\lambda y \beta(<x,y>), \lambda y \beta(<x+1,y>)]\}.$$

The principle implies classically the comprehension scheme [66]. It is *not* implied by the latter except when A in prenex form contains no universal set quantifier followed by an existential one. For independence [51], for exception [61]. However, this axiom of choice holds in *ramified* classical analysis, the analogue to 1.5, and thus 4.12 is finitistically reducible to 4.11.

4.13. The formal rules of *intuitionistic* logic are used instead of

[45] More precisely, to the system 4.13, 4.2211 resp. based on the formal rules of *intuitionistic* logic.

classical logic, the unrestricted axiom of choice, and any of the following schemata (for all A):

4.131. $\rightarrow\rightarrow \wedge x[A(x) \vee \rightarrow A(x)]$ or $[\wedge x \rightarrow\rightarrow A(x)] \rightarrow \rightarrow\rightarrow \wedge x A(x)$ or $\rightarrow[\rightarrow \wedge x A(x) \wedge \wedge x \rightarrow\rightarrow A(x)]$, [46], [66].

N.B. The axioms 4.131 are not valid for the interpretation of the logical connectives in 2.31. [Also one may add the continuity axioms (2.511).]

4.131 is formally a subsystem of 4.12. But 3.31 applied to analysis reduces 4.11 to 4.131. The reduction to 4.11 of the continuity axioms in addition to 4.131 is in [8] and alternatively in 4.223 below.

4.2. Now various formal extensions of the *bar theorem* 2.6231 are used instead of the comprehension scheme. They are auxiliaries toward the main result of this section.

4.21. First, systems with *classical* logic.

4.211. Add 2.6231 with *classical* logic. The derivation from the comprehension principle is standard, for the converse cf. [41].

4.212. First extend the language of 4.11 by adding symbols for functions of finite type, formally exactly as in [34] (cf. 2.22). Let α be a variable of type τ^0, β a variable over finite sequences of type τ object (by some suitable coding of finite sequences), u a type τ variable. The formal generalization of 2.6231 is

4.2121. $\wedge\alpha \vee x R(\alpha, x) \wedge \wedge xy\alpha\beta\{[R(\alpha, x) \wedge x < y \cdot \rightarrow R(\alpha, y)] \wedge [R(\alpha, x) \rightarrow Q(\bar{\alpha}x)] \wedge [\wedge u Q(\beta * u) \cdot \rightarrow Q(\beta)]\} \cdot \rightarrow \wedge\beta Q(\beta)$.

To get a model for this in classical analysis, one first interprets the higher type variables by means of neighborhood functions, as in 2.422. Since, here, $\wedge\alpha \vee x R(\alpha, x)$ expresses that a partial ordering of *functions* is well founded, the obvious proof of (the translation of) 4.2121 (in terms of neighborhood functions) needs the principle of dependent choices. So, as in 4.12, one uses the model of 4.1 provided by *ramified* classical analysis where an explicit well ordering of functions can be defined, as on page 105.

The converse is more surprising. The full axiom of choice (4.12) follows from 4.2121 by classical logic and the axiom of choice restricted to quantifier free A as in 2.3332. Further, 4.2121 need only be applied to purely existential formulae R. (This restriction will be used immediately.) [41.]

4.22. Now, *intuitionistic* logic, and quantifier free systems resp.

4.221. Intuitionistic logic of order ω, primitive recursive functions of order ω [34] (cf. 2.23), the axiom of choice, 4.2121, and the con-

tinuity axioms for objects b of type $0^{(\tau^0)}$, α, β of type τ^0

$$\wedge\alpha\vee x\wedge\beta[(\wedge z < x)(\alpha z = \beta z) \to b\alpha = b\beta].$$

4.2211. (For completeness:) 4.221 and continuity for *all* formulae A

$$\wedge\alpha\vee x A(\alpha, x) \to \wedge\alpha\vee x\vee y\wedge\beta[(\wedge z < y)(\alpha z = \beta z) \to A(\beta, x)].$$

4.221 is reduced to classical analysis because all the axioms are satisfied by the model of 4.212. The converse, and the reduction of 4.2211, are given in

4.222. *A quantifier free system of finite types:* The system of 2.23, that is, of [34], is extended by the following definitional scheme (bar recursion of finite type). Let c be a variable over finite sequences of type τ objects, $lh(c)$ its length, Y (the principal variable) of type $0^{(\tau^0)}$, $[c]$ the sequence (type τ^0 object) defined by the condition that its first $lh(c)$ members are c, the rest some fixed constant, and, the types of the variables a, b, are chosen so as to be coherent. For all types considered constants θ and axioms

$$\begin{aligned}\theta(a, b, c; Y) &= a(c) \text{ if } Y([c]) < lh(c);\\ &= b(\lambda u\theta(a, b, c * u, Y), c) \text{ otherwise;}\end{aligned}$$

are added, where u is a type τ variable.

These axioms are formally derivable from 4.221 and hence reducible to classical analysis.

4.2221. (An auxiliary system!) Add to 4.221 the schema 4.2121 with $\wedge\alpha\rightarrow\rightarrow\vee x R(\alpha, x)$ instead of $\wedge\alpha\vee x R(\alpha, x)$ for *purely existential* R, that is, $R(\alpha, x)$ has the form $\vee b^{\tau} R_1(\alpha, x, b)$, with quantifier-free R_1.

Main Result. 4.2221 *reduces to* 4.222 *when the interpretation of* [34], that is, 2.321, *is applied.*

More precisely, if A is a formula of classical analysis, A_1 the interpretation by 2.31 of the negative form of A as defined in 2.3322, then A can be proved from 4.12 if and only if A_1 can be proved in 4.222.

Corollary. Classical analysis reduces to 4.222, that is, there is a finitist consistency proof of classical analysis relative to 4.222.

For, by 4.212 (converse), classical analysis is essentially a subsystem of 4.2121 with purely existential R. By 3.31, this is reduced to intuitionistic logic and negative forms of the axioms, that is, 4.2221 instead of 4.221. But, by 2.322, on the interpretation 2.321,

4.2221 is equivalent to 4.2121 itself. This reduction by easy stages with interesting lemmas was given in [41] after a more complicated proof in [66].

4.223. If one wants to include 4.2211 one has to use the modified neighborhood functions of 2.422.

4.3. *For what notions of function of finite type is bar recursion* (4.222) *valid?* Not much seems to be known constructively without additional hypotheses. If *Church's thesis* is assumed the following cases (cf. 2.73) are settled:

4.31. Bar recursion at type 0, that is, for c in 4.222 ranging over finite sequences of numbers (type 0 objects), is false for effective operations.

4.32. For continuous functionals we have

4.321. (even without Church's thesis) bar recursion at types 0 and 1 can be proved, for example, in 2.622, and θ has a *recursive* neighborhood function,

4.322. bar recursion at type 2 is false if both Church's thesis *and* the axiom 2.521 are assumed.

N.B. the contrast with the classical theory of neighborhood functions (4.2121); in the proof of 4.322, Church's thesis distinguishes between, and 2.521 connects free choice sequences and constructive functions. It is open whether the thesis can be replaced by (2.531 and 2.511).

4.33. For the schemata of [44], a *definition* of θ can be written down, but there are no evident axioms for functionals of finite type to prove that θ is everywhere defined.

So the question 4.3 is wide open.

4.4. *Discussion.* Practically speaking, the results 4.31 and 4.32 are useful because, at the present time, we have no evident axioms for constructive functions of finite type which are not satisfied by the particular kinds of functionals considered in 4.3. But their significance for the proof theoretic study of classical analysis is strictly limited, even if we restrict ourselves to the particular idea of [46] which underlies the present section, namely this. The interpretation 2.32 can be formally applied to 4.131, which is interreducible with classical analysis; and one then asks: what plausible definition schemta are needed to satisfy (the interpretation of) the axioms 4.131?

We consider particularly the continuous case 4.32.

4.41. Continuity properties are *not* formally necessary because, by 2.3331, the axioms 4.131 are satisfied if arbitrary nonconstructive functionals are used.

4.42. More interesting, the analysis of [66, page 17] shows: it is sufficient (though not necessary) to solve the following

Problem. Consider the finite type structure over N (the set of natural numbers) and classes C_τ (of operations of each type) satisfying the following conditions: (a) They are closed under the primitive recursive operations of 2.23. (b) For C of type σ^0, Y of type $0^{(\sigma^0)}$, A, B, D with types that make sense: There should be functions B and C of A, D, Y in $\bigcup_\tau C_\tau$ which solve the functional equations

$$A(YC, B) = C(YC) \qquad \text{and} \qquad DC = B[C(YC)].$$

One way of satisfying (b) is this: $\bigcup_\tau C_\tau$ should contain the θ of 4.222 for all possible types, but, by 4.41, it is not the only way. A purely abstract study of the problem may be fruitful.

4.43. Finally, tricky definitions of classes of functionals aside (which satisfy 4.3), how promising is the general idea above in the light of present knowledge? That is, are the (negative) formal theorems of classical analysis valid on the interpretation 2.32 for the *standard* notion of functional of finite type?

4.431. Note first, that on 2.523, *negative* formulae have the same meaning whether type 0^0 variables are interpreted as constructive functions or as free choice sequences. So, if the interpretation 2.32 is faithful to 2.523, it is not reasonable to impose continuity requirements on the functionals used *unless free choice sequences are to be involved in proofs even of formulae not containing free choice variables* (cf. pages 133–134).

4.432. Next, consider 2.32 applied to the simplest nontrivial case of 4.131, where $A(x)$ is $\vee y B(x, y)$ with quantifier-free B: $\neg\neg \wedge x[\vee y B(x, y) \vee \neg \vee y B(x, y)]$, that is, $\neg\neg \wedge x \vee y \wedge z[B(x, y) \vee \neg B(x, z)]$. (This is intuitionistically equivalent, by use of the axiom of choice, to 4.11 with purely existential A.) On 2.32, writing $C(x, y, z)$ for $B(x, y) \vee \neg B(x, z)$

$$\wedge x \vee y \wedge z C(x, y, z) \qquad \text{means } (Ea)(xz)C(x, ax, z)$$

$$\neg \wedge x \vee y \wedge z C(x, y, z) \qquad \text{means } (Ex^2z^2)(a)C[x^2a, a(x^2a), z^2a].$$

If this is to be *established* without dubious hypotheses on con-

structive functions a, x^2 and z^2 must be defined for free choice sequences, that is, have strong continuity properties.

4.4321. $\rightarrow\rightarrow \wedge x \vee y \wedge z C(x, y, z)$ means $(x^2 z^2)(Ea)C[x^2 a, a(x^2 a), z^2 a]$; but if 4.4321 is to be established without dubious hypotheses it must *not* be *assumed* that x^2 and z^2 are necessarily defined for all free choice sequences. With this assumption, 4.4321 can be proved (even for arithmetic B), for example, in 2.622 by [66], but not without: specifically, by [46], 6.2 and 7.2, not for the extension of Church's thesis in [44].

5. ABOUT FOUNDATIONAL PROBLEMS

We now review briefly the relevance of the technical results of Sections 1–4 to the general problem on page 95.

5.1. (Hilbert's version of) *Formalism.* Recall the introduction and Section 3. The phenomena of reasoning are represented by combinatorial relations between formal expressions (words or formulae), so to speak, their outward and visible signs. Assertions which present themselves to us as being about abstract situations, are analyzed in terms of formal derivability of their representations. Only the manipulation of the formal expressions is counted as immediate experience. The fact that at least a partial representation, that is, a *descriptive* scheme, can be obtained in this way by use of simple mechanical rules is the discovery of formalization. These rules represent the principles of reasoning (in the domain of mathematics considered). The mere existence of a formalization which generates all (or some) accepted assertions does not explain or, as one says: justify the principles, in particular, it does not explain their application inside or outside mathematics (cf. page 150). Hilbert's program of consistency proofs or, more generally, interpretations gives a precise and coherent scheme for such an explanation in terms of a *reduction* to certain principles of reasoning, namely combinatorial proofs. These are a minimum used in all scientific reasoning (cf. footnote 31 for their relation to a mechanistic conception).

Modulo the open questions of 3.6 concerning combinatorial proofs, the reduction *cannot* be carried out for a well-established part of current mathematics, namely first order classical or intuitionistic arithmetic. And, for a wider version of Hilbert's program (pre-

dicative proofs), it cannot be carried out for current intuitionistic analysis and, a fortiori, not for classical analysis (2.63, 3.5).

For the areas where it can be carried out (3.3232) it provides a remarkable separation between (a) the reliability of combinatorial conclusions in reasoning which presents itself to us as being about certain abstract set theoretic or constructive objects, and (b) the abstract existential assumptions made in such reasoning. For wider areas this kind of separation is not possible (1.711).

The machinery of formal systems plays not only the basic role (formalization) described above, but a high degree of technical development of this machinery is used in an essential way. In particular, rules are needed in place of axioms for formalizing the concept of finitist proof; and this formalization, in turn, is needed for establishing the limitations on Hilbert's original program.

For reference below, note three things.

5.11. Gödel's incompleteness theorem is of central importance for the formalist conception.

5.12. Doctrinaire formalism. Why do we not regard the reasoning of classical number theory as simply unjustified, instead of regarding 3.4 as a proof of the limitations of Hilbert's program. Here it is to be remembered that the purpose of theory is not only to organize a limited set of data (here: baby arithmetic) or even to interpret and correct them, but, above all, *to extend the range of experience which can be theoretically understood.* From this point of view Hilbert's program, by itself, was always defective because it could justify principles, but not explain the choice of rules.

5.13. The refutation of formalism illustrates the well-known fact that one has to go beyond narrow experience (for example, of school mathematics for which Hilbert's program can be carried out remarkably well) to get decisive results on a theory. This is only reasonable because a theory would not be expected to survive if it were in conflict with very crude experience![46]

5.2. *Realism* and set theory. Here the emphasis is not on the process of reasoning, but on the results and, in particular, on the objects about which assertions are made. In consequence, if formalization is used as the descriptive scheme for reasoning, the justification of the rules consists in showing that the conclusions

[46] Galileo's first theory of falling bodies (velocity proportional to the distance fallen) did not survive long because according to it a body at rest would never start to fall ($s = Ae^{Bt}$) and this conflicts with very crude observation!

are *valid*, for example, 1.831, 1.835 for pure logic. While in 1.8 the rules themselves were not problematic, in the untutored reasoning about sets they were, and led to contradictions. Closer examination of what objects sets are (cumulative type structure) explains the restrictions on the rules. This is why 1.2 was regarded as a main result. As a theory, set theory is extremely attractive because it has few primitives, and other mathematical objects (natural numbers, real numbers, probabilities) are built up naturally: if it was an achievement to build up the physical world from a 100-odd atoms, how much more striking to build up the world of mathematics from two primitives: Furthermore, the laws for these primitives are elegant and surprisingly clear (1.67).

More generally, the realist conception is certainly very close to the way a good deal of mathematics presents itself to us, and (this was of course its prime reason) it explains the objectivity of mathematics, that is, agreement on results, by its being about external objects with which we are in some kind of contact. As has been pointed out in [33, 35], there is considerable similarity in the methods of acquiring knowledge in elementary mathematics and physics. Also, it would be agreed that the realist assumption of external mathematical objects is not more *dubious* than that of physical objects.

The *weakness* lies elsewhere. Basically, the trouble seems to be that the realist view has never been developed far enough to be put to a real test, in sharp contrast to (Hilbert's) formalism; cf. footnote 13. More specifically, it will be very interesting to see what considerations will settle the question of measurable ordinals (1.72, 1.733). I do not know a formulation of the realist view, for which experience establishes that there are infinite sets (in the sense of set theory), let alone inaccessible ones. Concerning the relation to physical objects it is certainly unusual in physics to regard the existence of objects as established because we can think of them, especially in advanced physics, and, by 5.13, one should perhaps not restrict oneself to the elementary case.

The important connection between set theoretic and arithmetic statements discovered in [32] (1.711) also differs in a fundamental respect from so-called observable consequences of physical hypotheses. Thus, if arithmetic statements are correlated with observable consequences [33], in some cases (on the basis of the known theory) the set theoretic hypotheses have no observable consequences

(1.535, 1.65) among them the true axiom of choice and the dubious continuum hypothesis. And in others, for example, existence of inaccessible ordinals, rival hypotheses are not in conflict, but one leaves a particular arithmetic question undecided while the other does not. By 5.12, this may be in support of the latter; but it seems to be unusual in the natural sciences to have rival theories where one *knows* that they will not conflict on observable consequences.

Perhaps one has stressed too much how "daring" and problematic the realist view is, instead of developing it so as to permit some conclusive decision. In the light of what has happened in the case of formalism a decision may be possible. For comparison note

5.21. Gödel's incompleteness theorem does not conflict with the realist view. In fact, if the subject matter is regarded as objective, one has no right to expect an intelligible theory, least of all a complete one in the form of a formal system! Further one has no reason to expect that the primitive objects can be grasped clearly.

5.22. While the *particular* realist conception formulated in the set theory of Zermelo-Frankel extends the range of theoretical understanding, for example, to establishing the consistency of the rules of classical analysis (and more), the realist view *tout court* does not: recall the remark above on whether there are infinitely large sets. In other words, it is not clear whether, on this view *as presently developed*, more or less formal reasoning is justified than on 5.1.

5.3. *Idealism and Intuitionistic Mathematics.* The idealist conception is probably the most commonplace one at the present time: mathematics is a free mental activity. As such it does not exclude the existence of mathematical objects external to us; they would have the same structure as the ideas involved, and so mathematics would be about them too. Certainly, at an elementary level, one does not ask oneself whether, for example, simple arithmetic statements are about concrete realizations (finite sets of things), one's ideas of such configurations, or about some abstract entities. Further, even if one admits abstract objects external to us, at *a certain stage* it would be quite in accordance with scientific practice to ignore them if the available information is dubious and confusing. This is done by the realist, too, who certainly (properly) ignores questions about the organization of the brain without denying that

this may be involved in a more delicate study of our relation to these external objects.

In this connection, as least as far as questions of *evidence* are concerned, even the particular *restriction* of the idealist conception discussed in Section 2, may recommend itself. Here (solipsism) one confines oneself to thoughts which are again about the mathematical activity itself and not about external objects. After all, in the process of reasoning as it presents itself to us, we benefit from teaching up to a certain elementary level, but after that understanding seems to be very much every man's own business. And if understanding is regarded as something that happens here and now, one adopts in effect the crude solipsist position. More pedantically, one seeks a theory of evidence whose parameters are those of the solipsist position. In short, there seems to be nothing outlandish in the general position.

The *weakness* of the general idealist position, as in 5.2, is rather that it is not sharp enough. In particular, I know of no formulation which excludes taking the whole of formal set theory as laws of our *ideas* of collections existing independently of us (whether or not they do). In fact, here it would be natural to take set theory as a *proper part* of a more general idealistic conception which deals not only with properties that define collections, but more general properties and to consider their laws (logic). For such a logic of not everywhere defined properties, the law of the excluded middle would not be valid. The notion of *collecting property* could then be defined, for example, as in [14], where a is collecting if for each t, $\vdash t \in a$ or $\vdash t \notin a$. This suggests the possibility of a *foundation* of the theory of sets in terms of more *general* notions, but to date the axioms for such a notion are weak.

5.31. More generally, *the primitive concepts of an idealist conception would not enter directly into mathematical practice, but would be used* for an analysis, that is, *for foundations.* See footnote 14, but also 3.31 where classical arithmetic concepts may be regarded as forming a *crude substructure* in which the delicate properties of (constructive) (Ex) and $\vee$ are suppressed.

5.311. A foundation along the lines of 5.31 is a kind of analogue to Hilbert's program, but with this essential difference: unlike the latter it would not reject any classical statements (as ideal 3.113: why should an idealistic conception reject ideal elements?) but

rather study statements about ideas which themselves are not necessarily about objects external to us. The execution of this kind of program may permit the outright *definition* of the (logical and non-logical) primitives of classical mathematics; just as the primitives of arithmetic are defined in set theory. (Outright, and not only contextually, in contrast to the elimination of free choice sequences in 2.523 or 2.622.)

5.32. Concerning the sharper version of idealism (solipsism) developed in intuitionistic mathematics, the main results of Section 2, summarized in 2.63, show that the *currently* justified formal rules cover roughly the part of analysis done at the turn of the century. Comparison with set theory is difficult because both the analysis of the basic notions of intuitionism along 2.2 and the theory of constructions of higher type in 2.8 are still quite crude. Not only technically, but also conceptually since the notion of set in the sense of the cumulative theory of types is certainly clearer than that of construction or proof.

5.321. It is not yet excluded that the bulk of classical mathematics can be interpreted in suitably sophisticated versions of 2.2 or 2.8. This would correspond to Hilbert's program with: intuitionistic proofs replacing combinatorial ones. As pointed out in 3.12(c), the interpretation problem is here more appropriate than the consistency problem since presumably at least all statements $\forall\alpha\exists x A(\alpha, x)$ for elementary A would be regarded as significant. A whole section (Section 4) was devoted to proof theory of impredicative analysis in view of the possibility just mentioned. Here, again, it is to be stressed that the proposal is *not to use the intuitionistic theories for mathematical practice, but for its foundation;* both conceptually, and for a technical analysis of the power of particular axiomatic theories. This technical use has been shown already both in standard proof theory of arithmetic (3.241) and, somewhat less directly (cf. footnote 9), of set theory.

5.322. There is work in the literature in a direction opposite to 5.321, namely to try and understand intuitionistic practice in terms of set theoretic concepts. For elementary parts, cf. [11]; for more highly developed areas, the results of [8] and the elimination results of Section 2 (2.6241) are important. Here abstract theories which use the concept of *proof*, as in extensions of 2.21, present difficulties. But while this work is often useful or even

essential for understanding the *formal* character of problems in intuitionistic mathematics, it has no direct interest for the idealist conception of mathematics.

5.323. Gödel's incompleteness theorem does not conflict directly with it, because there is no assumption that the laws of thought are mechanical (in the sense of: recursive). Cf. footnote 18 and 2.72.

5.324. The idealist conception is not antiphysical. In fact, there seems to be no reason why the facts of mathematical experience might not lead to general conclusions about physical laws (footnote 30).

5.4. *Idealism: problems.* From the discussion above (5.3, 5.323, 5.324) it seems *difficult to refute the idealist position by actual conflict with experience, or support it by establishing such conflict for its rival views.* While such simple criteria of success have been sufficient to decide between proposed theories in many areas of the natural sciences, different criteria may have to be applied here. (Or, in hackneyed language, the "operational significance" of the idealist position must be sought elsewhere.)

Briefly, cf. page 98 of the introduction, what characterizes the difference between e.g. the idealist and realist view is *what aspects of* (crude) *experience are regarded as significant and suitable for study.* Thus, in particular, the realist would hardly deny the fact of a great deal of agreement on various (intensional) properties of *proof* as in footnote 15: rather, he would regard it as insignificant for mathematics, and (probably) as not capable of precise theoretical treatment of any kind. So, a basic step towards a decision on the idealist position is this:

5.41. Can one develop a theory of such intensional properties? One would see the strength of the idealist position in its furnishing notions and laws in terms of which a wider area of experience can be interpreted (just as in 5.12). Some case of this kind has been made above (5.321) for the intuitionistic concepts. Evidently, not all experience that presents itself to us is suitable: just as in the natural sciences many (perhaps most!) aspects of experience, though perfectly objective, are taken as physically insignificant, and one does not try to set up a theory to explain them. For example, experience of particular, so-called accidental, facts. Therefore an essential step in the development (of the idealist position) is to select an area of experience and set oneself specific

tasks: just as for progress in pure mathematics it is essential to have mathematical problems (only now the mathematical formulation of the task is *part* of the solution, not *given* with the problem).

5.411. One such task, likely to be fruitful for the development of the idealist view, is the interpretation problem for classical mathematics (5.31 or 5.321).

5.412. But, perhaps, a more basic one is this, suggested by 3.4 (despite the still fragmentary character of this work): *to set up a theory for the detailed structure of proofs, not only for the class of provable theorems,* in short, characterizations of informal notions of proof. Some essential difficulties are mentioned in 3.6, but autonomous progressions [28] are certainly a useful intermediate step, at least for some of these informal notions. Experience suggests that there are relatively few such notions that are, as one says, *basic* in our mathematical experience, so that this area lends itself to theoretical analysis. At any rate (so it seems to me) *the essence of the idealistic conception consists precisely in the assumption that this impression is correct.*

5.42. Granted all this, two questions arise: is mathematics needed in this subject? Is a (primarily formal) mathematical study promising?

5.421. The *need* for an advanced mathematical treatment seems clear because the properties of proofs which are significant in reasoning, are complicated: so one needs highly developed technical machinery even to state *approximate* laws.

5.422. It would, perhaps, be merely frivolous to study extremely primitive elements of reasoning, like recognition of symbols, without regard for the physical processes involved. But, by 5.3 (page 187), this criticism should not apply to the areas of reasoning here considered. After all, a good deal of physical theory of matter in bulk is quite independent of understanding the atomic structure of matter (in fact, the theory would never have come about if one had first waited for such understanding): only relatively crude qualitative physical experience was used to good effect in early physical theory. An analogue to this is needed in the case of reasoning if a primarily mathematical analysis is to be promising.

These lectures were prepared during a stay at the Institute for Advanced Study at Princeton, for which I am much indebted to Professors Gödel and Oppenheimer. It is a pleasure to acknowledge that my conversations with Professor Gödel (during this stay and

on previous occasions) have helped me much, not least, to see some fruitful consequences of the general philosophic points made in his [33, 34, 35]. References in the text give a good idea of these consequences just because the views are sharp, and not merely suggestive in a literary way. But, I hope, also those parts of the lectures which are quite independent or, as in Sections 2 and 5, contrary to the general points mentioned, have been sharpened by the conversations.

REFERENCES

Whenever possible, books and monographs are cited instead of research papers; original articles only if they seem still the best exposition. No reference is made to abstracts which are adequately summarized in the text.

Books and Monographs

1. P. Bernays and A. A. Fraenkel, *Axiomatic set theory*, Amsterdam (1958) North Holland Publishing Co.
2. A. Church, The calculi of λ-conversion, *Ann. of Math. Studies*, **6**, Princeton (1951).
3. H. B. Curry and R. Feys, *Combinatory logic*, Amsterdam (1958) North Holland Publishing Co.
4. P. A. M. Dirac, *The principles of quantum mechanics*, Oxford, England (1947) Oxford University Press.
5. K. Gödel, The consistency of the axiom of choice and the generalized continuum-hypothesis with the axioms of set theory, *Ann. of Math. Studies*, **3**, Princeton (1951).
6. D. Hilbert and P. Bernays, *Grundlagen der Mathematik*, Vol. 2, Berlin (1939) Springer.
7. S. C. Kleene, *Introduction to metamathematics*, Princeton, N.J. (1952) Van Nostrand.
8. S. C. Kleene, *Foundations of intuitionistic mathematics*, Amsterdam (1965) North Holland Publishing Co.
9. R. Montague, D. Scott, and A. Tarski, *An axiomatic approach to set theory*, Amsterdam (1965) North Holland Publishing Co., to appear.
10. W. V. O. Quine, *Set theory and its logic*, Cambridge, Mass. (1963) Harvard University Press.
11. H. Rasiowa and R. Sikorski, *The mathematics of metamathematics*, Warsaw (1963) Pánstwowe Wydawnictwo Nankowe.
12. A. Robinson, *Introduction to model theory and to the metamathematics of algebra*, Amsterdam (1963) North Holland Publishing Co.
13. R. M. Smullyan, Theory of formal systems, *Ann. of Math. Studies*, **47**, Princeton (1961).

14. K. Schütte, *Beweistheorie*, Berlin (1960) Springer.
15. A. Tarski and J. C. C. McKinsey, *A decision method for elementary algebra and geometry*, Berkeley and Los Angeles (1951) Univ. of Calif. Press.
16. A. Tarski, A. Mostowski, and R. M. Robinson, *Undecidable theories*, Amsterdam (1953) North Holland Publishing Co.

Articles and Abstracts

17. W. Ackerman, Begründung des "tertium non datur," *Math. Ann.*, **93** (1924) 1–36.
18. W. Ackermann, Zur Axiomatik der Mengenlehre, *Math. Ann.*, **131** (1956) 336–345.
19. P. Bernays, Zur Frage der Unendlichkeitsschemata in der axiomatischen Mengenlehre, pp. 3–49 in *Essays on the foundations of mathematics*, Jerusalem (1961) Magnes Press.
20. L. J. B. Brouwer, Über Definitionsbereiche von Funktionen, *Math. Ann.*, **97** (1927) 60–76.
21. P. J. Cohen, A minimal model for set theory, *Bull. Amer. Math. Soc.*, **69** (1963) 537–540.
22. P. J. Cohen, The independence of the continuum hypothesis I, II, *Proc. Nat. Acad. Sc.*, **50** (1963) 1143–1148; **51** (1964) 105–110.
23. J. N. Crossley, R. Parikh, On isomorphisms of recursive well orderings, *J.S.L.*, **28** (1963) 308.
24. B. Dreben, P. Andrews, S. Aanderaa, False lemmas in Herbrand, *Bull. Amer. Math. Soc.*, **69** (1963) 699–706.
25. A. Ehrenfeucht, Separable theories, *Bull. Acad. Pol. Sc.*, **9** (1961) 521–524.
26. S. Feferman, Arithmetisation of metamathematics in a general setting, *F.M.*, **49** (1960) 36–92.
27. S. Feferman, Classification of recursive functions by means of hierarchies, *Trans. Amer. Math. Soc.*, **104** (1962) 101–122.
28. S. Feferman, Transfinite recursive progressions of axiomatic theories, *J.S.L.*, **27** (1962) 259–316.
29. S. Feferman, Systems of predicative analysis, *J.S.L.*, **29** (1964) 1–30.
30. R. O. Gandy, On a problem of Kleene's, *Bull. Amer. Math. Soc.*, **66** (1960) 501–503.
31. G. Gentzen, Untersuchungen über das logische Schliessen, *Math. Zeitschrift*, **39** (1934) 176–210.
32. K. Gödel, Uber formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, *Monatshefte Math. Physik*, **38** (1931) 173–198.
33. K. Gödel, Russell's mathematical logic, pp. 123–153 in *The philosophy of Bertrand Russell*, Evanston and Chicago (1944) Northwestern University Press.
34. K. Gödel, Über eine bisher noch nicht benützte Erweiterung des finiten Standpunktes, *Dialectica*, **12** (1958) 280–287.
35. K. Gödel, What is Cantor's continuum problem, pp. 258–273 in P. Benacerraf and H. Putnam, *Philosophy of mathematics*, Englewood Cliffs, N.J. (1964) Prentice-Hall.
36. W. Hanf, Incompactness in languages with infinitely long expressions, *FM*, **53** (1964) 309–324.

37. W. Hanf and D. Scott, Classifying inaccessible cardinals, *Notices of Amer. Math. Soc.*, **8** (1961) 445.
38. J. Herbrand, Recherches sur la théorie de la démonstration, *Travaux Soc. Sc.*, **33** (1930) Warsaw.
39. A. Heyting, Die formalen Regeln der intuitionistischen Logik, *Preuss. Akad. Wiss.* (1930) 42–56, 57–71, 158–169.
40. G. Higman, Subgroups of finitely presented groups, *Proc. Roy. Soc.*, **A262** (1961) 455–474.
41. W. A. Howard, Bar induction, bar recursion, and the Σ_1^1 – axiom of choice, *J.S.L.*, **28** (1963) 303.
42. H. J. Keisler and A. Tarski, From accessible to inaccessible cardinals, *FM*, **53** (1964) 225–308.
43. S. C. Kleene, Hierarchies of number theoretic predicates, *Bull. Am. Math. Soc.*, **61** (1955) 193–213.
44. S. C. Kleene, Recursive functionals and quantifiers of finite types, *Trans. Amer. Math. Soc.*, **91** (1959) 1–52.
45. S. C. Kleene, Disjunction and existence under implication in elementary intuitionistic formalisms, *J.S.L.*, **27** (1962) 11–18.
46. G. Kreisel, Interpretation of classical analysis by means of constructive functionals of finite type, pp. 101–128 in *Constructivity in mathematics*, Amsterdam (1959) North Holland Publishing Co.
47. G. Kreisel, A remark on free choice sequences and the topological completeness proofs, *J.S.L.*, **23** (1958) 369–388.
48. G. Kreisel, On weak completeness of intuitionistic predicate logic, *J.S.L.*, **27** (1962) 139–158.
49. S. Kripke, The undecidability of monadic modal quantification theory, *Zeitschrift f. math. Logik und Grundlagenforschung*, **8** (1962) 113–116.
50. A. Levy, On Ackermann's set theory, *J.S.L.*, **24** (1959) 154–166.
51. A. Levy, Independence results in set theory by Cohen's method, I–IV, *Notices Amer. Math. Soc.*, **10** (1963) 592–593.
52. M. H. Loeb, Solution of a problem of Leon Henkin, *J.S.L.*, **20** (1955) 115–118.
53. R. Montague, Semantical closure and non-finite axiomatisability, pp. 45–69 in *Infinitistic methods*, Warsaw (1961) Państwowe Wydawnictwo Nankowe.
54. R. Montague, Set theory and higher order logic, pp. 131–148 in *Formal systems and recursive functions*, Amsterdam (1964) North Holland Publishing Co.
55. M. Morley, On theories categorical in uncountable powers, *Proc. Nat. Acad. Sc.*, **49** (1963) 213–216.
56. A. Mostowski, Representability of sets in formal systems, *Proc. Symp. Pure Maths.*, **5** (1962) 29–48.
57. C. D. Parsons, The Ω-consistency of ramified analysis, *Archiv. f. math. Logik und Grundlagenforschung*, **6** (1962) 30–34.
58. H. Putnam, On hierarchies and systems of notations, *Proc. Amer. Math. Soc.*, **15** (1964) 44–50.
59. A. Robinson, On generalized limits and linear functionals, *Pac. J. Math.*, **14** (1964) 269–283.
60. F. Rowbottom, Thesis, University of Wisconsin, Madison (1964).

61. Y. Sampei, On the uniformization of the complement of an analytic set, *Comment. Math. Univ. St. Paul*, **10** (1962) 57–62.
62. D. Scott, Measurable cardinals and constructible sets, *Bull. Ac. Pol. Sc.*, **9** (1961) 521–524.
63. J. R. Shoenfield, The problem of predicativity, pp. 132–139 in *Essays on the foundations of mathematics*, Jerusalem (1961) Magnes Press.
64. R. Solovay, Independence results in the theory of cardinals, *Notices of the Am. Math. Soc.*, **10** (1963) 595.
65. C. Spector, Recursive well-orderings, *J.S.L.*, **20** (1955) 151–163.
66. C. Spector, Provably recursive functionals of analysis: a consistency proof of analysis by an extension of principles formulated in current intuitionistic mathematics, *Proc. Symp. Pure Maths.*, **5** (1962) 1–27.
67. W. W. Tait, Infinitely long terms of transfinite types, pp. 176–185 in *Formal systems and recursive functions*, Amsterdam (1964) North Holland Publ. Co.
68. W. W. Tait, The ϵ-substitution method, *J.S.L.*, to appear.
69. W. W. Tait, The no-counterexample interpretation, *Archiv für math. Logik und Grundlagenforschung*, to appear.
70. G. Takeuti, Construction of the set theory from the theory of ordinal numbers, *J. Math. Soc. Japan*, **6** (1954) 196–220.
71. G. Takeuti, On the theory of ordinal numbers, *J. Math. Soc. Japan*, **9** (1957) 93–113.
72. G. Takeuti, On the fundamental conjecture of GLC, GLC (I–V), *J. Math. Soc. Japan*, **10** (1958) 121–134, and earlier papers quoted there.
73. J. Ax, S. Kochen, Diophantine problems over local fields, *Amer. J. Math.*, to appear.
74. A. Lévy, Measurable cardinals and the continuum hypothesis, *Notices Amer. Math. Soc.*, **11** (1964) 769–770.
75. R. Solovay, The measure problem, *Annals of Mathematics*, to appear.

4

Some Recent Advances and Current Problems in Number Theory

Paul Erdös

The subject of number theory is very extensive and has intimate links with other branches of mathematics. Analysis has been applied with great success to various problems of number theory for 200 years and there is every reason to expect that such applications will continue. Algebraic methods have also been applied to number theory and have in turn developed out of number theory. Recently algebraic geometry and probability theory have been applied with success to problems which previously seemed intractable. In this chapter I clearly cannot hope even to attempt to give a complete survey of recent developments in number theory, and in quite a few of its branches I am not particularly competent to do so—for example, in the branches involving algebraic geometry. My paper will be highly subjective; I shall write mainly about questions which have interested me personally, and I certainly do not wish to suggest that any problems and results which I omit to mention are less important or interesting than the ones I shall write about a great deal. For instance, I overemphasize problems on primes and problems of a combinatorial type; also, of course, I overemphasize my own work. I shall not write much about Waring's

problem since it has been dealt with in recent books [1]; I shall omit the geometry of numbers, and also Diophantine approximation since I recently wrote about this subject (see my forthcoming paper in *Compositio Math.*). The same fate will overtake many applications of probability to number theory, but several survey articles have appeared recently on this subject (some of them written by me) and there is also a recent book by Kubilius and a forthcoming book by Rényi and myself [2]. Most of the questions with which I shall deal will have a combinatorial flavor or will relate to primes (or both); these are the subjects which have interested me most for the last thirty-three years. To quote from the introduction to the well-known and excellent book of Hardy and Wright: "I cannot fail completely in making the paper interesting, since the subject is so attractive that this would need extravagant incompetence."

There will be some overlap between this paper and my recent paper "On unsolved problems" [2a].

I wish to thank my friends Davenport, Schinzel, and Turan for their valuable assistance.

1. First, I shall discuss problems and results on the distribution of prime numbers (the letters p, q will denote primes throughout).

Denote by $\pi(x)$ the number of primes not exceeding x and by c, c_1, . . . absolute constants, not always the same. The Prime Number Theorem states that

$$\lim_{x=\infty} \frac{\pi(x)}{\frac{x}{\log x}} = 1. \tag{1}$$

(1) was first proved in 1896 by Hadamard and de la Vallée Poussin. In 1948 Selberg and I obtained [3] an elementary proof of (1). Our starting point was the following remarkable formula of Selberg, which he proved in an elementary way:

If $\vartheta(x) = \sum_{p \le x} \log p$ then

$$\vartheta(x) \log x + \sum_{p<x} \log p \, \vartheta\left(\frac{x}{p}\right) = 2x \log x + O(x). \tag{2}$$

We then proved by elementary arguments that if $1 < p_1 < p_2 <$

. . . is any sequence of real numbers which satisfies (2) and further satisfies

$$\vartheta(x) > ax, \qquad \sum_{p<x} \frac{\log p}{p} = \log x + O(1), \tag{3}$$

then

$$\vartheta(x) = x + o(x). \tag{4}$$

This is well known to be equivalent to (1).

Beurling [4] gave the following interesting generalization of the prime number theorem. Let $1 < p_1 < p_2 < \cdots$ be any sequence of real numbers, which will be called generalized primes. Denote by $N(x)$ the number of solutions of

$$\prod_i p_i^{\alpha_i} \leq x \qquad (\alpha_i = 0, 1, \ldots).$$

Assume that

$$N(x) = x + o\left(\frac{x}{(\log x)^{3/2}}\right). \tag{5}$$

Then it follows that

$$\sum_{p_i \leq x} 1 = (1 + o(1)) \frac{x}{\log x}. \tag{6}$$

Beurling also observed that (6) cannot be deduced from a weaker error term than that in (5). Selberg observed that one can make the deduction of (6) from (5) by our elementary method if in (5) a slightly better error term is assumed (unpublished). Nyman and Malliavin [4] sharpened Beurling's results in various ways. I later proved [3] that (2) alone implies (4) and Shapiro [3] proved that (2) with an error term $o(x \log x)$ instead of $O(x)$ also implies (4).

My deduction of (4) from (2) was based on the following Tauberian theorem, which seems of independent interest:

Suppose $a_k \geq 0$ and assume that $\left(\text{with } s_n = \sum_{k=1}^{n} a_k\right)$

$$\sum_{k=1}^{n} a_k(s_{n-k} + k) = n^2 + O(n). \tag{7}$$

Then

$$s_n = n + O(1). \tag{8}$$

My original proof was very complicated; Siegel simplified it (unpublished) and later Shapiro [3] simplified my proof considerably.

Bombieri and Wirsing [5] succeeded independently in proving in an elementary way that for every $k > 0$

$$\vartheta(x) = x + o\left(\frac{x}{(\log x)^k}\right). \tag{9}$$

This, as is well known, implies that

$$\pi(x) = \int_o^x \frac{dy}{\log y} + o\left(\frac{x}{(\log x)^k}\right).$$

This is a considerable advance on previous results of Van der Corput, Kuhn, Breusch, and others.

In my opinion the simplest deduction of (4) from (2) and (3) is that due to V. Nevanlinna [6], who somewhat simplifies the proof of Wright [1].

Put $\psi(x) = \sum_{n<x} \Lambda(n)$, $\Lambda(n) = \log n$ if $n = p^a$ and is 0 otherwise. Tchebicheff observed that

$$\sum_{n=1}^{x} \psi\left(\frac{x}{n}\right) = x \log x - x + o(x) \tag{10}$$

The proof of (10) is elementary; in fact (10) easily follows from a weak form of Stirling's formula.

It would be very desirable to deduce the prime number theorem from (10) as far as possible in an elementary way. Sharpening a previous result of Landau, Ingham [7] proved by using Wiener's theory that (10) implies $\psi(x) = x + o(x)$ which is equivalent to the prime number theorem. Recently Ingham and I proved the following theorem (our paper will appear in *Acta Arithmetica*): Let $1 < a_1 < a_2 \cdots$ be a sequence of real numbers with $\sum_i 1/a_i < \infty$. Assume that $f(x)$ is an increasing function for which

$$f(x) + \sum_i f\left(\frac{x}{a_i}\right) = x\left(1 + \sum_i \frac{1}{a_i}\right) + o(x). \tag{11}$$

(11) implies $f(x) = x + o(x)$ if and only if $\sum_i 1/(a_i)^s = 1$ has no root of the forms $1 + it$, $t \neq 0$. It is possible that if the a_i are integers this condition is always satisfied. The simplest case for which we cannot decide this question is $a_1 = 2$, $a_2 = 3$, $a_3 = 5$. Several related problems are discussed in our paper.

The sharpest estimation of $\vartheta(x)$ is at present

$$\vartheta(x) = x + O(x \exp(-(\log x)^{3/5-\epsilon})),$$

obtained by Korobov and Vinogradoff [8]. The Riemann Hypothesis would imply

$$\vartheta(x) = x + O(x^{1/2} \log x)$$

Now I go on to state some problems and results about the distribution of primes. Let $2 = p_1 < p_2 < \cdots$ be the sequence of consecutive primes. A well-known theorem of Tchebicheff states that $p_{n+1} < 2p_n$ for all n, and the prime number theorem implies that $p_{n+1}/p_n \to 1$. The sharpest upper bound for $p_{n+1} - p_n = d_n$ is due to Haneke [9]; he proved, sharpening previous results of Hoheisel, Haneke, Heilbronn, Ingham, and Min, that

$$d_n < p_n^{9/20+\epsilon}.$$

The Riemann Hypothesis would imply

$$d_n < p_n^{1/2+\epsilon}.$$

It has been conjectured that between two consecutive squares there is always a prime. This conjecture can probably not be deduced from the Riemann Hypothesis and seems to be very deep. Piltz conjectured that for every $\epsilon > 0$ and $n > n_0(\epsilon)$

$$d_n < n^{\epsilon}.$$

Cramer [10] conjectured that

$$\overline{\lim_{n=\infty}} \frac{d_n}{(\log n)^2} = 1. \tag{12}$$

Cramer was lead to his conjecture by probabilistic reasoning; the proof or disproof of (12) seems hopeless by the methods which are at our disposal at present. It has been known for a very long time

that $\overline{\lim}\, d_n = \infty$ (the $n-1$ integers $n!+2, n!+3, \ldots, n!+n$ are all composite).

Sharpening previous results of Backlund, Brauer-Zeitz, and Westzynthius, I proved [11] by using Brun's method that for infinitely many n

$$d_n > c \log n \log\log n/(\log\log\log n)^2. \tag{13}$$

Chang [11] succeeded by a simple further idea in dispensing with Brun's method in the proof of (13), and Rankin using an improvement of our method proved that for infinitely many n and every $\epsilon > 0$

$$d_n > (\tfrac{1}{3} - \epsilon) \log n \log\log n \log\log\log\log n/(\log\log\log n)^2. \tag{14}$$

(14) seems to be the natural boundary of our method; the only improvement of (14) in the last 26 years is due to Schönhage and Rankin, who replaced the constant $\frac{1}{3}$ by e^{γ} [11].

A well-known and probably very difficult conjecture on primes asserts that

$$\pi(x+y) \leq \pi(x) + \pi(y). \tag{15}$$

Hardy and Littlewood [12] proved by Brun's method that

$$\pi(x+y) - \pi(x) < \frac{cy}{\log y}. \tag{16}$$

As far as I know this is the only time Hardy and Littlewood used Brun's method. Selberg [12] improved (16) to

$$\pi(x+y) - \pi(x) < \frac{2y}{\log y} + O\left(\frac{y \log\log y}{(\log y)^2}\right). \tag{17}$$

It would be very important if one would replace 2 in (17) by a smaller constant, but this seems to be difficult. A slightly weaker conjecture than (15) states that, for every ϵ and $y > y_0(\epsilon)$,

$$\pi(x+y) - \pi(x) < \frac{(1+\epsilon)y}{\log y}.$$

Selberg's investigations [12] on the limits of the efficiency of the sieve methods indicate that (17) cannot be improved by Brun's method except possibly if very essential changes are made in the estimation of the error terms.

It seems likely that $d_n/\log n$ is everywhere dense in $(0, \infty)$. Ricci [13] and I proved independently that the limit points of $d_n/\log n$ form a set of positive Lebesgue measure, but the only known limit point of this set is ∞. In particular I do not know if $d_n/\log n$ has a rational limit point. It seems certain that $d_n/\log n$ has a continuous distribution function $\psi(t)$. In fact, Bombieri conjectures that $\psi(t) = 1 - e^{-t}$, in other words, the density of integers n for which $d_n < t \log n$ equals $1 - e^{-t}$. There does not seem to be much hope of attacking this conjecture at present, but later in this paper I will state a modification which can probably be settled.

Ricci [13] proved by Brun's method that the lower density of the integers n for which $d_n < \log n$ is positive, and in fact it is not hard to show that there is an $\epsilon > 0$ such that the lower density of the integers n for which $d_n < (1 - \epsilon) \log n$ is also positive. Unfortunately I cannot prove that the upper density of the integers n for which

$$d_n > \log n$$

is positive. It is not hard to deduce from Brun's method that

$$\Sigma' d_n > cx,$$

where the dash indicates that the summation is extended over those $p_n < x$ for which $d_n > \log n$. But unfortunately nothing can be deduced from this because of possible very large values of d_n. In fact I just observe to my annoyance that I cannot show that $d_n/\log n$ has at least one finite limit point greater than or equal to 1. One could give by Brun's method a rough estimation for a constant c so that $d_n/\log n$ certainly has a finite limit point $\geq c$, but as far as I know nobody has given an explicit value for c.

I proved [14] that

$$\underline{\lim}\, d_n/\log n < 1$$

(the prime number theorem immediately implies that the $\underline{\lim}$ is ≤ 1). Rankin [14] proved that the limit in question is $\leq \frac{57}{59}$; Ricci [13] showed that it is $\leq \frac{15}{16}$, and finally Bombieri proved that it is $\leq \frac{29}{32}$ (unpublished). There seems to be no doubt that the $\underline{\lim}$ in question is 0, but this seems very hard to prove; the well-known conjecture that there are infinitely many prime twins, that is, that $d_n = 2$ has infinitely many solutions, would of course imply this.

The sequence d_n, $n = 1, 2, \ldots$, behaves very irregularly.

Turán and I [15] proved that the inequalities $d_{n+1} > d_n$ and $d_{n+1} < d_n$ both have infinitely many solutions. We have not been able to prove that $d_n > d_{n+1} > d_{n+2}$ or $d_n < d_{n+1} < d_{n+2}$ have infinitely many solutions. In fact we cannot disprove the existence of an integer n_0 so that for every $k \geq 0$, $d_{n_0} + 2k > d_{n_0} + 2k + 1$ and $d_{n_0} + 2k + 1 < d_{n_0} + 2k + 2$. It is not known that $d_n = d_{n+1}$ has infinitely many solutions; Rényi and I [15] proved that the number of solutions of $d_n = d_{n+1}$, $1 \leq n \leq x$ is less than $cx/(\log x)^{1/2}$; very likely the true order is $cx/\log x$ but this seems difficult.

It is not difficult to show that for infinitely many indices n, $d_n > d_{n+1}$, $d_n > d_{n-1}$ and for infinitely many indices m, $d_m < d_{m+1}$, $d_m < d_{m-1}$.

Sierpinski [16] observed that

$$\overline{\lim_n} \min (d_n, d_{n+1}) = \infty. \tag{18}$$

It is perhaps surprising that though the proof of $\overline{\lim} d_n = \infty$ is trivial the proof of (18) is much more difficult.

Walfisz and Prachar [16] proved that the upper density of the integers n for which

$$\min (d_n, d_{n+1}, \ldots, d_{n+k}) < \epsilon \log n$$

tends to 0 for fixed k together with ϵ (I slightly modified their result). Also I observed [16] that for every $c_1 > 0$ there exists a $c_2 > 0$ so that there are at least $c_2 \log n$ consecutive values $d_k, \ldots, d_{k+r}$, $(k < n)$, all of which are $> c_1$, but I do not know if this holds for every c_1 and c_2. Rényi and I [15] also showed that

$$\frac{c_1 n}{\log n} < \sum_{k=1}^{n} \frac{1}{d_k} < \frac{c_2 n \log\log n}{\log n};$$

probably the upper bound is nearly best possible.

Prachar and I [17] proved that

$$c_1 (\log x)^2 < \sum_{p_k < x} \left| \frac{p_{k+1}}{k+1} - \frac{p_k}{k} \right| < c_2 (\log x)^2.$$

We further showed that if k_i is a subsequence of the integers for

which

$$\frac{p_{k_i}}{k_i} < \frac{p_{k_{i+1}}}{k_{i+1}},$$

then the density of this sequence is 0.

Put $p_k/k = u_k$. We further observed that the prime number theorem implies, for $k > k_0(\epsilon)$,

$$u_{[k(1+\epsilon)]} > u_k; \tag{19}$$

on the other hand to every l there are infinitely many values of k for which

$$u_k > u_{k+l}. \tag{20}$$

There is a big gap between (19) and (20) which we cannot fill. In our paper we state the following further problems on the structure of the sequence u_k:

There are only a finite number of values of k (possibly none) for which

$$\max_{1\le i\le k} u_{k-i} < u_k < \min_{1\le i\le \infty} u_{k+i}.$$

We easily show that the density of the integers k for which $u_k > u_{k+1}$ is positive. We cannot show that the same holds for the k for which $u_k < u_{k+1}$.

We do not know if $u_k < u_{k+1} < u_{k+2}$ or $u_k > u_{k+1} > u_{k+2}$ has infinitely many solutions.

Returning to the question of d_n, I may mention that by using Brun's method I proved [18] that

$$\overline{\lim} \min \frac{(d_n, d_{n+1})}{\log n} = \infty$$

but I cannot prove that

$$\underline{\lim} \max \frac{(d_n, d_{n+1})}{\log n} < 1 \text{ or } \overline{\lim} \min \frac{(d_n, d_{n+1}, d_{n+2})}{\log n} = \infty,$$

also I cannot prove that

$$\underline{\lim} \frac{d_n + \cdots + d_{n+k-1}}{k \log n} < 1 - c$$

where c does not depend on k.

It seems certain that the density of the indices n for which $d_n > d_{n+1}$ is $\frac{1}{2}$, but this seems very hard to prove. I proved [19] though that for a sufficiently small $\epsilon > 0$ the lower density of the indices n for which $d_n > (1 + \epsilon)d_{n+1}$ is positive, and the same result holds for $(1 + \epsilon)d_n < d_{n+1}$.

Define $n_1 < n_2 < \cdots$ as follows:

$$d_{n_i} > d_n \qquad \text{for all } n < n_i.$$

Very little is known about the sequence n_i, for example, I cannot prove that $n_{i+1} > n_i + 1$ for $i > i_0$. It is easy though to see that the density of the n_i is 0.

Cramér [10] proved, assuming the Riemann hypothesis, that

$$\sum_{n<x} d_n^2 < cx(\log x)^4.$$

Very probably

$$\sum_{n<x} d_n^2 < cx(\log x)^2,$$

but this seems hopeless. It may even be true that

$$\lim_{x=\infty} \frac{1}{(\log x)^2} \sum_{n=x} d_n^2 = c. \tag{21}$$

It is not hard to prove that the lower limit in (21) is positive.

Similar questions can be asked for other sequences of numbers. For example, let $s_1 < s_2 < \cdots$ be the sequence of squarefree numbers; it is well known and easy to prove that their density is $6/\pi^2$. I proved [20] that

$$\sum_{s_i<n} (s_{i+1} - s_i)^2 = c_2 n + o(n).$$

It seems very probable that for every $\alpha > 0$

$$\sum_{s_i<n} (s_{i+1} - s_i)^\alpha = c_\alpha n + o(n). \tag{22}$$

(22) if true must be very difficult, since it would imply $s_{i+1} - s_i = o(s_i^\epsilon)$ for every $\epsilon > 0$. My method breaks down for $\alpha > 2$ but it proves (22) for $\alpha \leq 2$. The best upper bound for $s_{i+1} - s_i$ is due to Richert [20] who proved (sharpening a previous result of

K. F. Roth)

$$s_{i+1} - s_i < cs_i^{3/6} \log s_i.$$

It is easy to prove [20] that

$$s_{i+1} - s_i > (1 + o(1)) \frac{\pi^2}{6} \log s_i / \log \log s_i, \tag{23}$$

but as far as I know nobody has succeeded in replacing $1 + o(1)$ by $1 + c$ in (23).

Denote by $Q(x)$ the number of squarefree integers not exceeding x. It is easy to prove that

$$Q(x) = \frac{6}{\pi^2} x + O(x^{1/2}), \tag{24}$$

and the prime number theorem gives $o(x^{1/2})$ in (24). One would expect the Riemann Hypothesis to give $o(x^{1/4+\epsilon})$, but it seems that one can only deduce $o(x^{3/6+\epsilon})$. The true order of magnitude of the error term in (24) is unknown.

One could try to generalize (22) as follows. Let $a_1 < a_2 < \cdots$ be an infinite sequence of integers satisfying $a_k/k^2 \to \infty$ and let $b_1 < b_2 < \cdots$ be the sequence of integers no one of which is a multiple of any a. Is it then true that

$$\sum_{b_i \leq n} (b_{i+1} - b_i)^2 = Cn + o(n)? \tag{25}$$

If $a_i = p_i^2$ we obtain (22). If instead of $a_k/k^2 \to \infty$ only $a_k < ck^2$ is assumed it is easy to see that (25) cannot hold; at present I cannot disprove that in this case

$$\sum_{b_i < n} (b_{i+1} - b_i)^2 < An$$

remains true for a suitable A.

In [14] I conjectured that if $1 = a_1 < a_2 < \cdots < a_{\varphi(n)} = n - 1$ are the integers relatively prime to n then

$$\sum_{i=1}^{\varphi(n)-1} (a_{i+1} - a_i)^2 < \frac{cn^2}{\varphi(n)}. \tag{26}$$

This conjecture seems to be an elementary version of (21) and should not be too difficult to prove.

Hooley [20] in fact proved that for every $\alpha < 2$

$$\sum_{i=1}^{\varphi(n)-1} (a_{i+1} - a_i)^\alpha < \frac{c_\alpha n^\alpha}{\varphi(n)^{\alpha-1}}$$

and

$$\sum_{i=1}^{\varphi(n)-1} (a_{i+1} - a_i)^2 < cn^2 \frac{\log\log n}{\varphi(n)}.$$

Put $n_k = 2, 3, \ldots p_k$. The $\varphi(n_k)$ integers relatively prime to n_k in the interval $(1, n_k)$ might be expected to show a somewhat similar behavior to the primes. Let

$$1 = a_1^{(k)} < a_2^{(k)} < \cdots < a_{\varphi(n_k)}^{(k)} = n_k - 1$$

be the integers relatively prime to n_k. Let us investigate to what extent this sequence satisfies the conjectures we stated about primes. First of all, it is not hard to deduce from Brun's method that there are constants c_1 and c_2 such that every interval of length $c_1(\log n)^{c_2}$ contains an integer relatively prime to n_k.

A theorem of Mertens implies that

$$\frac{n_k}{\varphi(n_k)} = \frac{(1 + o(1))e^{-\gamma}}{\log\log n_k} = \frac{(1 + o(1)e^{-\gamma}}{\log k}.$$

It is not hard to prove that (if $a_{i+1}^{(k)} - a_i^{(k)} = d_i^{(k)}$) the sequence

$$\frac{d_i^{(k)}}{\log k}$$

is everywhere dense in $(0, \infty)$; in other words to every ϵ and η there is a k_0 such that for $k > k_0$ every interval of length η in $(\epsilon, 1/\epsilon)$ contains a number of the form $d_i^{(k)}/\log k$. I have not been able to prove that

$$\frac{d_i^{(k)}}{\log k}$$

has a distribution function (the precise meaning of this statement is obvious and is left to the reader).

It seems probable that the number of integers $1 \leq s < \varphi(n_k)$ for which $d_{i+1}^{(k)} > d_i^{(k)}$, is $[\frac{1}{2} + o(1)]\varphi(n_k)$, but as far as I know this has

not been proved. I do not know the number of solutions of $d_{i+1}^{(k)} = d_i^{(k)}$, but it would be easy to obtain crude upper and lower bounds.

It is easy to see that for every t and all $k > k_0(t)$

$$d_i^{(k)} > d_{i+1}^{(k)} > \cdots > d_{i+t-1}^{(k)}$$

is solvable.

Sivasankaranarayana Pillai conjectured that

$$\sum_{\substack{p_n<x \\ n\equiv 0 \pmod 2}} d_n = [\tfrac{1}{2} + o(1)]x. \tag{27}$$

(27) seems very hard to prove; Brun's method easily gives

$$\sum_{\substack{p_n<x \\ n\equiv 0 \pmod 2}} d_n > cx.$$

One can also conjecture that

$$\sum_{i\equiv 0 \pmod 2} d_i^{(k)} = [\tfrac{1}{2} + o(1)]n_k, \tag{28}$$

but I have not been able to prove this.

Jacobsthal defines $g(n)$ to be the least integer such that among any $g(n)$ consecutive integers there is at least one relatively prime to n. Put

$$\max g(n) = C(r) + 1,$$

where the maximum is taken over all the integers n with $\nu(n) \leq r$ (where $\nu(n)$ denotes the number of distinct prime factors of n). We have

$$\frac{c_1 r(\log r)^2 \log\log\log r}{(\log\log r)^2} < C(r) < c_2 r^{c_3}. \tag{29}$$

The left side of (29) follows from (13) and the right side can be easily obtained by Brun's method. Jacobsthal conjectured that

$$C(r) < c_4 r^2. \tag{30}$$

The exponent in (29) can be reduced by Selberg's improvement of Brun's method, but (30) seems hopeless at present [21].

Now we discuss primes in arithmetic progressions. Dirichlet was the first to prove that every arithmetic progression $\{a + kd\}$ with $(a, d) = 1$ represents infinitely many primes. Many mathematicians attempted without success to find an elementary proof, but

finally Selberg [22] was successful. Denote by $\pi(a, d, x)$ the number of primes $\leq x$ of the form $a + kd$. The prime number theorem for arithmetic progressions states that

$$\pi(a, d, x) = [1 + o(1)]\frac{x}{\varphi(d) \log x} \qquad (d \text{ fixed}, x \to \infty). \tag{31}$$

It is not difficult to prove (31) by the method of Selberg and myself [22]. The generalized Riemann Hypothesis for L-functions would imply that

$$\pi(a, d, x) = \frac{1}{\varphi(d)} \int_2^x \frac{dy}{\log y} + O(x^{1/2} \log x)$$

uniformly in d, and also that the least prime $p(a, d)$ in $a + kd$ is less than $d^{2+\epsilon}$.

Linnik [23] proved without using any hypothesis that

$$p(a, d) < c_1 d^{c_2}.$$

Linnik's proof has been simplified first by Rodosskij and still further recently by Turán and Knapowski [22].

Turán [24] proved using the generalized Riemann Hypothesis that for all but $o[\varphi(d)]$ arithmetic progressions $a + kd$

$$p(a, d) < cd(\log d)^{2+\epsilon}. \tag{32}$$

Perhaps the exponent $2 + \epsilon$ can be replaced by $1 + \epsilon$ but this is very deep if true. I proved [24] using Brun's method that for every $c_1 > 0$

$$p(a, d) < c_1 \varphi(d) \log d$$

for at least $c_2\varphi(d)$ [where $c_2 = c_2(c_1)$] values of a. In the opposite direction, I could only show that there exists a constant c_3 and an infinite sequence $d_1 < d_2 < \cdots$ such that

$$p(a, d_i) > (1 + c_1)\varphi(d_i) \log d_i \tag{33}$$

for at least $c_4\varphi(d_i)$ values of a. There seems no doubt that this holds for all sufficiently large d, but I could not prove it. The proof of (33) used Brun's methods and thus strongly used special properties of primes. Perhaps the following general result holds: Let $a_1 < a_2 < \cdots$ be a sequence of integers for which

$$A(x) = \sum_{a_i \leq x} 1 = [1 + o(1)] \frac{x}{\log x}.$$

Denote by $f(a, d, n)$ the smallest $a_i \equiv d \pmod{n}$. Then there is an infinite sequence $n_1 < n_2 < \cdots$ so that, for at least $c_1 n_i$ values of d in $0 \leq d < n_i$,

$$f(a, d, n_i) > (1 + c_2) n_i \log n_i. \tag{34}$$

Perhaps (34) holds for all sufficiently large n.

Using the results of [24] Turán proved that for every irrational $\alpha > 1$ the sequence $p\alpha \pmod 1$ is uniformly distributed; later Vinogradov [24] proved this without any hypothesis, and by using his powerful methods of estimating trigonometric sums, he also obtains a fairly good estimation of the discrepancy of the sequence $p\alpha \pmod 1$. It follows easily from the uniformity of distribution that for every irrational $\alpha > 1$, $[n\alpha] = p$ has infinitely many solutions. As far as I know it is not known whether there are infinitely many primes p for which $[p\alpha] = q$.

Now I want to say something about the comparative theory of prime numbers; a subject recently developed by Turán and Knapowski [25]. The origin of this subject is to be found in the following conjecture of Tchebicheff: put

$$f(x) = \sum_p (-1)^{p-1/2} e^{-px}.$$

Then Tchebicheff stated that $f(x) \to -\infty$ as x tends to 0. This conjecture is still unproved and must be very deep since Hardy, Littlewood, and Landau [25] showed that it is equivalent to the Riemann Hypothesis for the L-function $1 - \frac{1}{3^s} + \frac{1}{5^s} + \frac{1}{7^s} + \cdots$ belonging to the modulus 4.

Tchebicheff stated that his conjecture implies a preponderance of the primes $\equiv 3 \pmod 4$ over those $\equiv 1 \pmod 4$. Littlewood [25] proved on the other hand that

$$\pi(1, 4, x) - \pi(3, 4, x)$$

changes sign infinitely often.

Turán and Knapowski [25] recently took up this subject and obtained a whole series of interesting results which seemed unattainable previously. I just state here a few of them and must refer to their joint papers and to their forthcoming book on this subject.

A modulus d is called good if the L-functions $L(s, \chi)$ belonging to this modulus d have no real root in the critical strip. The values $3 \leq d \leq 12$ are all good and possibly every modulus is good, but it is not even known that there are infinitely many good moduli.

1. If d is good and T is sufficiently large, then the interval[1] $(\log_3 T, T)$ always contains an x_1 and an x_2 for which for any l with $(l, d) = 1$ and $l \not\equiv 1 \pmod{d}$,

$$\pi(1, d, x_1) - \pi(l, d, x_1) > \frac{\sqrt{x_1}}{\log x_1} \log_5 x_1,$$

$$\pi(1, d, x_2) - \pi(l, d, x_2) < \frac{\sqrt{x_2}}{\log x_2} \log_5 x_2;$$

further $\pi(1, d, x) - \pi(l, d, x)$ has at least $c \log_4 T$ changes of sign in $(0, T)$.

2. For good d the interval $(\log_3 T, T)$ always contains an x_1 and x_2 for which

$$\pi(1, d, x_1) - \frac{\pi(x_1)}{\varphi(d)} > \frac{\sqrt{x_1}}{\log x_1} \log_5 x_1,$$

$$\pi(1, d, x_2) - \frac{\pi(x_2)}{\varphi(d)} < -\frac{\sqrt{x_2}}{\log x_2} \log_5 x_2.$$

3. If d is good and $T > T_0$ and l is a quadratic residue (mod d) then the interval $(T^{1/2}, T)$ contains values x_1 and x_2 for which

$$\pi(1, d, x_1) - \pi(l, d, x_1) > T^{1/2} \exp\left(-\frac{\log T \log_3 T}{\log_2 T}\right),$$

$$\pi(1, d, x_2) - \pi(l, d, x_2) < -T^{1/2} \exp\left(-\frac{\log T \log_3 T}{\log_2 T}\right).$$

4. If d is good and $T > T_0$ then for every two distinct values l_1 and l_2

$$\sum_{\substack{n \equiv l_1 \pmod{d} \\ n \leq x}} \Lambda(n) - \sum_{\substack{n \equiv l_2 \pmod{d} \\ n \leq x}} \Lambda(n)$$

changes sign in $(T, e^{2\sqrt{T}})$.

[1] We write $\log_2$ for log log, and so on.

5. Knapowski [26] proved that for sufficiently large T

$$\pi(x) - \int_2^x \frac{dy}{\log y}$$

changes sign at least $c \log_4 T$ times in $(0, T)$. (Riemann conjectured that

$$\pi(x) < \int_2^x \frac{dy}{\log y}$$

for all x and Littlewood disproved this conjecture by showing that

$$\pi(x) - \int_2^x \frac{dy}{\log y}$$

$L_i(x)$ changes sign infinitely often [26].)

Knapowski and Turán use the new and surprising inequalities of Turán which he developed in several of his papers and in his book [25]. A new English edition of the book will soon appear and will contain many interesting problems and new results. The inequalities are analytic in nature but can also be considered as part of the theory of Diophantine approximation, and in a certain sense they can be considered as generalizations of Dirichlet's theorem. Here I want to state only two problems in this theory, on which I also worked.

Let $z_1 = 1$ and $|z_i| \leq 1$ for $2 \leq i \leq n$. Put $s_k = \sum_{i=1}^{n} z_i^k$. Turán conjectured that there exists an absolute constant c such that, for all n and all choices of the z's,

$$\max_{1 \leq k \leq n} |s_k| \geq c. \tag{35}$$

Atkinson [27] recently proved this conjecture, and in an unpublished manuscript he showed that c can be chosen to be $\frac{1}{8}$. Turán further conjectured that to every ϵ there is an n_0 such that for $n > n_0$

$$\max_{1 \leq k \leq n} |s_k| > 1 - \epsilon.$$

I observed [25] that there is a constant $c > 0$ and a sequence with $z_1 = 1$, $|z_i| \leq 1$ for $2 \leq i \leq n$ such that

$$\max_{2 \leq k \leq n+1} |s_k| < \frac{1}{(1+c)^n}. \tag{36}$$

The contrast between (35) and (36) is striking. I was unable to decide the existence of a sequence with $|z_i| \geq 1$ for $1 \leq i \leq n$ which satisfies (36).

Very recently Turán told me the following conjecture: Assume that the infinite sequence s_k, $1 \leq k < \infty$, contains infinitely many consecutive $n - 1$-tuples which are all 0. Then (essentially)

$$z_j = c^{2\pi ij/n}, \qquad 1 \leq j \leq n. \tag{37}$$

Perhaps (37) can be deduced if we know only that the sequence s_k contains two consecutive $n - 1$-tuples which are 0.

Before I leave the subject of prime numbers I would like to call attention to some related questions: E. Jabotinsky and I and independently and simultaneously V. Gardiner, R. Lazarus, N. Metropolis, and S. Ulam considered a modification of the sieve of Eratostenes and we were lead to several interesting questions, but for this I must refer to our papers on this subject [28].

Very interesting questions are raised in a paper by Hawkins on the so-called random sieve [29]; since this is perhaps not very well known I give the necessary definitions. We define a "random" sequence $a_i(t)$ as follows: Put $a_1 = 2$, and cross out each integer $3, 4, \ldots$ with probability $\frac{1}{2}$. Let a_2 be the first integer which has not been crossed out. Then cross out each of the integers $a_2 + 1$, $a_2 + 2, \ldots$ with probability $1/a_2$ and let a_3 be the first integer not crossed out, then cross out each of the integers $a_3 + 1, \ldots$ with probability $1/a_3$, and so on. Thus we obtain the "random" sequence $a_i = a_i(t)$, and Hawkins conjectured that for almost all t $a_i/(i \log i) \to 1$, but as far as I know this has never been satisfactorily proved [29].

Finally, I would like to call attention to the following result: Let $p_1 = 3$, $p_2 = 5$, and let p_k be the smallest prime for which

$$p_k \not\equiv 1 (\mathrm{mod}\ p_i), \qquad 1 \leq i < k.$$

Then I prove that

$$\lim_{k=\infty} \frac{p_k}{k \log k \log \log k} = 1. \tag{38}$$

The proof of (38) uses Tauberian arguments which are simpler than those used in the elementary proof of the prime number theorem [30].

6. Now I discuss some results in the arithmetic theory of polynomials. Let $f(x) = a_0 x^n + \cdots + a_n$ be an irreducible poly-

nomial with integral coefficients. Denote by $\nu(p)$ the number of solutions of the congruence

$$f(x) \equiv 0 \pmod{p}.$$

The prime ideal theorem [22] states that

$$\sum_{p<x} \nu(p) = (1 + o(1)) \frac{x}{\log x}. \tag{39}$$

Shapiro [22] proved (39) by the method of Selberg and myself in an elementary way. The proof is elementary in the sense that it does not use function theory, but as in all the other proofs of (39) he has to use algebraic number theory, that is, the theory of ideals. It would of course be interesting to prove (39) without the use of ideal theory, but perhaps this is not possible. I often tried without success to prove without using ideal theory that

$$\sum_{\nu(p)>0} 1/p = \infty.$$

Knapowski and Turán (unpublished; see [25]) proved the following theorem: Suppose $T > T_0(f)$. Then there are four numbers u_1, u_2, u_3, u_4 satisfying

$$\log^3 T \leq u_2 \exp(-8(\log u_2)^{5/6}) \leq u_1 \leq u_2 \leq T,$$

$$\log^3 T \leq u_4 \exp(-8(\log u_4)^{5/6}) \leq u_3 \leq u_4 \leq T,$$

for which

$$\sum_{u_1 \leq p \leq u_2} \nu(p) - \int_{u_1}^{u_2} \frac{du}{\log u} > \frac{u_2^{1/2}}{\log u_2},$$

$$\sum_{u_3 \leq p \leq u_4} \nu(p) - \int_{u_3}^{u_4} \frac{du}{\log u} < -\frac{u_4^{1/2}}{\log u_4}.$$

This theorem is new even for the case $f(x) = x$.

Denote by $\nu(m)$ the number of distinct prime factors of m. The prime ideal theorem immediately implies that

$$\sum_{n=1}^{x} \nu[f(n)] = [1 + o(1)]x \log\log x.$$

Turán [31] proved the following surprising result: Let $h(n)$ tend to infinity together with n as slowly as we please; then the density of the integers n for which the inequality

$$\log\log n - h(n)(\log\log n)^{\frac{1}{2}} < \nu[f(n)] < \log\log n + h(n)(\log\log n)^{\frac{1}{2}}$$

does not hold is 0. The special case $f(x) = x$ is a classical result of Hardy and Ramanujan [31].

Halberstam [31] proved that the density of integers n for which

$$\nu[f(n)] < \log\log n + c(\log\log n)^{\frac{1}{2}}$$

equals $(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{c} e^{-x^2/2}\,dx$. The special case $f(x) = x$ is contained in a theorem of Kac and myself.

I proved [31] that the number of primes $p \leq x$ for which

$$(1 - \epsilon)\log\log p < \nu(p-1) < (1+\epsilon)\log\log p \tag{40}$$

is not satisfied is $o(x/\log x)$. Halberstam [31] proved the following very much more general and more precise result. Suppose $f(x) \neq cx$. Then the number of primes $p \leq x$ for which

$$\nu[f(p)] < \log\log p + c(\log\log p)^{\frac{1}{2}}$$

equals

$$[1 + o(1)](2\pi)^{-\frac{1}{2}}\left(\int_{-\infty}^{c} e^{-y^2/2}\,dy\right)\frac{x}{\log x}.$$

Denote by $d(n)$ the number of divisors of n. Titchmarsh [31] proved that

$$\frac{c_2 x}{(\log x)^{\frac{1}{2}}} < \sum_{p<x} d(p-1) < c_1 x.$$

I proved [31] using (40) that

$$\sum_{p<x} d(p-1) > \frac{x}{\log x}\, 2^{(1-\epsilon)\log\log x}, \tag{41}$$

and Haselgrove [32] proved that

$$\sum_{p<x} d(p-1) > \frac{cx}{\log\log x}.$$

Finally, Linnik [31], using his powerful new dispersion method, proved that

$$\sum_{p<x} d(p-1) = (1 + o(1)) \frac{315\zeta(3)}{2\pi^4} x + o(x).$$

Van der Corput [32] proved that

$$c_1 x \log x < \sum_{n=1}^{x} d[f(n)] < c_2 x (\log x)^{\alpha}.$$

I proved [32] that

$$\sum_{n=1}^{x} d[f(n)] < c_3 x \log x \tag{42}$$

The proof is elementary but not simple. Very likely

$$\sum_{n=1}^{x} d[f(n)] = cx \log x + o(x \log x). \tag{43}$$

If true (43) must be very hard to prove, since the prime factors greater than x make the sharp estimation of the sum (43) very difficult. The constant c in (43) will perhaps depend on the polynomial $f(x)$. Bellmann and Shapiro [32] proved (43) if $f(x)$ is of degree 2, and in this case $c = 2$. Recently Hooley [32] proved that if $f(x)$ is of degree 2 then

$$\sum_{n=1}^{x} d[f(n)] = 2x \log x + O(x^{\alpha}), \qquad \alpha < 1.$$

Using Brun's method and the one with which I proved (42), I can show that

$$\sum_{p<x} d[f(p)] < c_4 x.$$

Perhaps if $f(x) \neq cx$ one can show by Linnik's method that

$$\Sigma\, d[f(p)] > c_5 x.$$

Denote by $P(n)$ the greatest prime factor of n. Tchebicheff [33] proved that

$$\lim_{x=\infty} P\Big[\prod_{n=1}^{x} (1 + n^2)\Big]/x = \infty.$$

Nagell and Ricci [33] proved that if $f(x)$ is of degree greater than 1 then

$$P\Big[\prod_{n=1}^{x} f(n)\Big] > cx \log x,$$

and I [33] proved that

$$P\Big[\prod_{n=1}^{x} f(n)\Big] > c_1 x\,(\log x)^{\log_2 x}.$$

By more complicated methods I can prove that

$$P\Big[\prod_{n=1}^{x} f(n)\Big] > c_3 x \exp\,(\log x)^{c_4}. \tag{44}$$

I never published the proof of (44), which is fairly complicated. The proof could be simplified a great deal if I could prove the following purely combinatorial theorem [33]. To every c_1 there exists a c_2 so that if $A_1, \ldots, A_l$, where $l = [c_2{}^k]$, are sets each having at most k elements, then there are c_1 of them $A_{i_1}, \ldots, A_{i_{c_1}}$ which have pairwise the same intersection. Rado and I [33] proved this with $k!(c_1 - 1)^k$ instead of $[c_2{}^k]$. (44) seems to be the natural boundary of my method. Very likely

$$P\Big[\prod_{n=1}^{x} f(n)\Big] > x^{1+d}, \tag{45}$$

but this seems very difficult. (44) would follow easily if we could prove that the number of integers $n \leq x$ for which all prime factors of $f(n)$ are $\leq x$ is greater than cx, but this has not even proved for $f(x) = 1 + x^2$.

It seems probable in fact that

$$P\Big[\prod_{n=1}^{x} f(n)\Big] > cx^k$$

when $f(x)$ is a polynomial of degree k.

A well-known result of Pólya [34] states that if the degree of $f(x)$ is > 1 then

$$\lim_{n=\infty} P[f(n)] = \infty. \tag{46}$$

If $f(x) = 1 + x^2$ Mahler and Chowla [34] (independently) proved that

$$P[f(n)] > c \log \log n. \tag{47}$$

(47) is certainly very far from being best possible, but I do not know of any reasonable upper bound for $P[f(n)]$ which is valid for infinitely many n. [Added in proof: Schinzel just informed me that he showed that for every t there are infinitely many n for which $P(n^2 + t) < \exp(c \log n/\log \log \log n)$.]

Another result of Pólya [34], related to (41), states that if $p_1, \ldots, p_k$ is any finite set of primes and $a_1 < a_2 < \cdots$ is the set of all integers composed of the p's, then $a_{i+1} - a_i$ tends to infinity. This was improved by Siegel [34] to

$$a_{i+1} - a_i > a_i^{1-\epsilon} \tag{48}$$

for every $\epsilon > 0$ if $i > i_0(\epsilon)$. It is easy to see that if $k > 1$ then

$$\frac{a_{i+1}}{a_i} \to 1 \tag{49}$$

as $i \to \infty$. There is a gap between (48) and (49) which as far as I know has not yet been filled.

Here I would like to mention a problem of Wintner which he communicated to me orally. Does there exist an infinite sequence of primes $p_1 < p_2 < \cdots$ such that if $a_1 < a_2 < \cdots$ is the set of all the integers composed of the p's then

$$\lim_{i=\infty} (a_{i+1} - a_i) = \infty ? \tag{50}$$

It seems certain that such a sequence exists, but I was unable to prove this.

One final result about greatest prime factors. It is not difficult to prove that for every n

$$P(2^n - 1) > n.$$

Schinzel [34] recently proved that for $n > 12$

$$P(2^n - 1) > 2n. \tag{51}$$

The proof of (51) is surprisingly complicated. Very likely

$$\lim_{n=\infty} P(2^n - 1)/n = \infty.$$

As far as I know $P(n! + 1)$ has not yet been investigated.

The polynomial $x^2 + x + 2$ is even for all integral values of x, thus it cannot represent any odd prime. A well known elementary theorem states that if $f(x)$ is of degree k and $f(x) \equiv O(\text{mod } n)$ for all integral values of x then $n \mid k!$. One could conjecture that $f(n)$ represents infinitely many integers of the form $a \cdot p$ where $a \mid k!$. As already stated Dirichlet proved this for polynomials of degree 1 in 1837. But the conjecture has never been proved even for a single polynomial of degree greater than 1. The only result for expressions of degree greater than 1 is due to I. I. Pjatezkij-Schapiro [35]. He proved that the number of primes of the form $[n^c]$ in $1 \leq n \leq x$ is $(1 + o(1))x/(1 + c) \log x$ if $1 \leq c \leq \frac{12}{11}$.

The result very likely holds for all nonintegral $c \geq 1$.

Heilbronn [36] proved using Brun's method that the number of integers $n \leq x$ for which $f(n)$ is a prime is less than $cn/\log n$; also it follows from Brun's method that there is an absolute constant c_1, depending only on the degree of $f(x)$, such that $f(n)$ represents infinitely many integers having fewer than c_1 prime factors.

Another conjecture states that $f(x)$ represents infinitely many integers of the form aQ, where $a \mid k!$ and Q is squarefree. It is well known and easy to prove that $f(x)$ represents infinitely many kth power-free integers, and in fact the density of the integers n for which $f(n)$ is kth power-free is positive (k being the degree of $f(x)$).

I proved [37] that if $k > 2$ then $f(x)$ represents infinitely many $(k-1)$-th power-free integers, the only exception being that if $k = 2^l$ then it may happen that $f(n) \equiv O(\text{mod } 2^{l-1})$ for all n, but then $f(x)$ represents infinitely many integers of the form $2^{l-1}Q$ where Q is odd and $(k-1)$-th power-free. The proof is fairly complicated. We would expect that the density of the integers for which $f(n)$ is $(k-1)$-th power-free is positive, but I could not prove this. I could prove nothing about the representation of $(k-2)$-th power-free numbers, for example, I cannot show that $n^4 + 2$ represents infinitely many squarefree numbers.

As far as I know the question whether $2^n \pm 1$ represents infinitely many kth power-free integers, is intractable at present, and the same is true for $n! \pm 1$.

7. Now I want to discuss a set of problems which could be said to belong to combinatorial number theory, that is, the questions have both number-theoretic and combinatorial character. These problems perhaps do not all have great importance but they are very

close to my heart (or rather I should say to my brain) since two of my main interests are number theory and combinatorial analysis.

I will start with Van der Waerden's theorem [38]. This asserts that if one splits the integers into two classes in any way then at least one of them contains an arbitrary long arithmetic progression. We shall be more concerned here with the finite form of this theorem, also proved by Van der Waerden: For every k there exists a smallest integer $f(k)$ such that if we split the integers $1 \leq t \leq f(k)$ into two classes then at least one of them contains an arithmetic progression of k terms. The upper bound obtained for $f(k)$ is enormously large; the reason for this is that all proofs use a double induction. Denote by $f(k, l)$ the smallest integer such that if we split the integers $1 \leq t \leq f(k, l)$ into l classes then at least one of them contains an arithmetic progression of k terms. The induction is carried out with respect to k and l and so gives a very poor estimation for $f(k, l)$ and in particular for $f(k) = f(k, 2)$. At least I believe that the estimation is very bad, though no one succeeded in obtaining any better one. Rado and I [38] obtained the first nontrivial lower bound for $f(k)$, by proving that $f(k) > ((k-1)2^k)^{1/2}$. The proof is based on the following simple consideration: The total number of ways of splitting n integers into two classes is clearly 2^n, and the number of splittings such that one of the two sets contains a given arithmetic progression of k terms is easily seen to be 2^{n-k+1}, and since there are fewer than n^2 arithmetic progressions all of whose terms are $\leq n$ we obtain that the total number of ways of splitting the integers $1 \leq t \leq n$ so that one of the sets will contain an arithmetic progression of k terms is at most $n^2 2^{n-k+1}$. This is less than 2^n if $n < 2^{(k-1)/2}$, whence $f(k) \geq 2^{(k-1)/2}$, and a more careful estimation of the number of arithmetic progressions gives $f(k) > [(k-1)2^k]^{1/2}$.

W. Schmidt [38] obtained by a difficult and ingenious improvement of our method

$$f(k) > 2^{k-c(k \log k)^{1/2}}, \tag{52}$$

and this is the best known lower bound for $f(k)$ up to the present time.

Using Van der Waerden's theorem, A. Brauer [39] proved that to every k there is a $p_0(k)$ so that if $p > p_0(k)$ then p has k consecutive quadratic residues and also k consecutive quadratic nonresidues (in fact Brauer proved a somewhat more general theorem).

Probably the right order of magnitude of $p_0(k)$ is $\exp ck$. It can be deduced from the results of A. Weil on congruences in two variables that $p_0(k) < \exp ck$.

In the same way as for the well-known theorem of Ramsey, define $g(k, l)$ as the smallest integer for which if we split the integers $1 \leq i \leq g(k, l)$ into two classes then either the first class contains an arithmetic progression of k terms or the second an arithmetic progression of l terms. If $k \leq l$ then clearly $f(k) \leq g(k, l) \leq f(l)$, but I do not know of any nontrivial estimation of $g(k, l)$; in particular it would be interesting to have upper and lower estimations of $g(3, l)$.

Let $h(n)$ be an arbitrary number-theoretic function which takes on the values $+1$ and -1. Van der Waerden's theorem asserts that for every k there is an arithmetic progression for which $h(a) = h(a + d) = \cdots = h[a + (k - 1)d]$. For a long time I conjectured that for every c there exist a d and an m such that

$$\left| \sum_{k=1}^{m} h(kd) \right| > c; \tag{53}$$

more precisely, perhaps there exists a constant c such that for every function $h(n)$ and every x there exist d and m with $md \leq x$ such that

$$\left| \sum_{k=1}^{m} h(kd) \right| > c_1 \log x. \tag{54}$$

It is easy to see that (54) if true is best possible. (53) requires much less than Van der Waerden's theorem but the arithmetic progressions are much more restricted. K. F. Roth recently proved (to appear in *Acta Arithmetica*) that there exist $a > 0$, $d < n^{1/2}$, $a + md < n$ for which

$$\left| \sum_{k=0}^{m} h(a + kd) \right| > cn^{1/4}. \tag{54'}$$

Roth in fact proves a more general theorem. In conversation Roth raised the question whether if we drop the condition $d < n^{1/2}$, then $cn^{1/4}$ can perhaps be replaced by $n^{1/2-\epsilon}$. I showed by probabilistic reasoning that (54') is false in general with $cn^{1/2}$ when c is sufficiently large. It is probably false with $cn^{1/2}$ for every $c > 0$ if $n > n_0(c)$, but I have not been able to show this.

Assume now that $h(n) = \pm 1$ is multiplicative, that is, $h(a \cdot b) = h(a) \cdot h(b)$. Then (53) would imply that

$$\left| \sum_{k=1}^{m} h(k) \right| \tag{55}$$

is unbounded. (55) seems quite difficult [38]. If $h(p^\alpha) = (-1)^\alpha$ we obtain Liouville's function $\lambda(n)$ and (55) is well known in this case, in fact

$$\sum_{k=1}^{n} \lambda(k) \neq o(n^{\frac{1}{2}-\epsilon}).$$

An interesting and beautiful conjecture on multiplicative functions $h(n) = \pm 1$ states that

$$\lim_{x=\infty} \frac{1}{x} \sum_{n=1}^{x} h(n) \tag{56}$$

always exists and that the limit of (56) is 0 if and only if

$$\sum_{h(p)=-1} \frac{1}{p} = \infty. \tag{57}$$

If (57) does not hold it is easy to see that the limit (56) exists and is different from 0. The conjecture (56) does not seem easy; it certainly contains the prime number theorem, for if $h(n) = \lambda(n)$, (56) is well known to be equivalent to the prime number theorem.

Wintner observed that if we only assume $|h(n)| = 1$ then (56) does not have to hold. Rényi recently observed that $h(n) = e^{i \log n}$ provides a simple counterexample. [Added in proof: Wirsing just informs me that he proved (56).]

If $h(n)^k = 1$ for some k, (56) probably remains true.

Another variant of Van der Waerden's theorem is the following: Define $f(k, c)$ $(0 < c \leq 1)$ as the smallest integer such that for every $h(n) = \pm 1$ there exists an arithmetic progression

$$0 < a < a + d < \cdots < a + (k-1)d < f(k, c)$$

for which

$$\left| \sum_{l=0}^{k-1} h(a + ld) \right| \geq ck.$$

Clearly $f(k, 1) = f(k)$. Using the same method as that which Rado and I used, I showed [38] that for every $c > 0$

$$f(k, c) > (1 + \alpha_c)^k,$$

where $\alpha_c \to 0$ as $c \to 0$ and $\alpha_c \to \sqrt{2} - 1$ as $c \to 1$ (perhaps the method of Schmidt would allow one to prove that $\alpha_c \to 1$ as $c \to 1$, but this has not been done). I would expect that for every $c < 1$

$$f(k, c) < (1 + \alpha_c')^k.$$

I am doubtful whether the same inequality holds for $f(k)$ (that is, for $c = 1$). Possibly

$$\lim_{k=\infty} [f(k, c)]^{1/k} = \alpha(c), \qquad 0 < c < 1.$$

The problem of obtaining a good upper bound for $f(k)$ led Turán and myself [40] to the following question: Let $1 \leq a_1 \leq \cdots \leq a_l \leq n$ and assume that the sequence $\{a_i\}$ does not contain an arithmetic progression of k terms. Put

$$\max l = \gamma_k(n).$$

If we could show that, for a certain n, $\gamma_k(n) < n/2$, we would immediately obtain $f(k) \leq n$. Unfortunately this has been shown only for $k = 3$, $n = 20$. In our paper [40] we only obtain crude inequalities for $\gamma_3(n)$, and Szekeres conjectured that $\gamma_k(n) = o(n)$ and that $\gamma_k(n) < n^{1-\epsilon_k}$. Behrend [40] observed that $\lim_{k=\infty} \gamma_k(n)/n = c_k$ exists; he further showed that either all $c_k = 0$ or $\lim_{k=\infty} c_k = 1$. Salem and Spencer [40] disproved $\gamma_k(n) < n^{1-\epsilon_k}$, in fact they showed that $\gamma_3(n) > n^{1-c/\log\log n}$, and Behrend showed [40] that

$$\gamma_3(n) > n^{1-c/\sqrt{\log n}}. \tag{58}$$

This is still the best known lower bound for $\gamma_3(n)$. Rankin [40] improved (58) for $k > 3$.

Roth [40] proved $\gamma_3(n) = o(n)$, in fact he showed that

$$\gamma_3(n) < \frac{cn}{\log\log n}.$$

Unfortunately Roth's method does not seem to give $\gamma_4(n) = o(n)$. It would be of great interest if one could prove that $\gamma_k(n) < \pi(n)$ for every k if $n > n_0(k)$ since this would imply that for every k there

are k primes in an arithmetic progression. Sevedinskij observed that $23143 + l \cdot 30030$ is a prime for $0 \leq l \leq 11$. Chowla [40] proved that there are infinitely many triplets of primes in arithmetic progression. His proof does not use $\gamma_k(n)$, but runs as follows. As is well known Vinogradoff proved that every sufficiently large odd number is the sum of three primes, and Van der Corput, Esterman, and Tchudakoff proved using Vinogradoff's method that the number of even numbers not exceeding x which are not of the form $p + q$ ($p \neq q$) is $o(x/(\log x)^k)$ for every k. Thus for infinitely many primes r, $p + q = 2r$ is solvable, hence there are infinitely many triplets of primes in arithmetic progression. This proof no longer works for quadruplets, and I do not see how a proof could be obtained except through an estimation of $\gamma_k(n)$. If $\gamma_k(n) < \pi(n)$ could be proved, then the fact that the primes contain arbitrarily long arithmetic progressions would be deduced just from the fact that the primes are numerous and would not use any special properties of the primes. This method is sometimes successful, for example, I proved [41] in this way that to every k there is an n_k such that $n_k = p^2 - q^2$ has more than k solutions.

Denote by $f_k(n)$ the number of solutions of

$$\sum_{i=1}^{k} p_i^k = n.$$

I proved in [41] that $\limsup f_2(n) = \infty$; the proof used special properties of primes, but it could be modified so as to use only $\pi(n) > n/(\log n)^k$. I can also prove that $\limsup_{n=\infty} f_3(n) = \infty$ (unpublished), and the proof of this seems to need some special properties of the primes. I can prove nothing for $k > 3$.

Now I have to break my word given in the introduction, since I have to mention a conjecture of Hardy and Littlewood which is of importance in Waring's problem: Denote by $\psi_k(n)$ the number of solutions in positive integers of

$$\sum_{i=1}^{k} x_i^k = n.$$

The famous K-hypothesis of Hardy and Littlewood asserts that $\psi_k(n) = o(n^\epsilon)$ for every $\epsilon > 0$.

It is well known that this would imply $G(k) < ck$, in other words

every sufficiently large integer is the sum of at most ck positive integral kth powers (more precisely the K-hypothesis would imply $G(k) \leq 4k$ for all k, and $G(k) \leq 2k + 1$ if k is not a power of 2).

For $k = 2$ the K-hypothesis is well known to hold. For $k = 3$ Mahler [42] disproved the K-hypothesis; he showed by an identity that

$$\psi_3(n^{12}) > cn. \tag{59}$$

As far as I know $\overline{\lim}_{n=\infty} \psi_3(n)/\log n$ has not been determined. The K-hypothesis is probably wrong for $k > 3$ too but this has never been proved.

For the applications to Waring's problem it would suffice to show that for every $\epsilon > 0$:

$$\sum_{k=1}^{x} \psi_k(n)^2 = o(x^{1+\epsilon}). \tag{60}$$

(60) is probably true but this has never been proved.

Chowla, Pillai, and I [42] proved that for every k and infinitely many n

$$\psi_k(n) > \exp(c_k \log n/\log\log n). \tag{61}$$

(61) is of course not enough to disprove the K-hypothesis.

Let A be a sequence of integers of positive density. Denote by $\psi_k(A; n)$ the number of solutions of

$$\sum_{j=1}^{k} a_{i_j}{}^k = n.$$

I can prove (unpublished) that

$$\limsup \psi_k(A; n) = \infty. \tag{62}$$

More generally if c_1 and c_2 are given and $n > n_0(c_1, c_2)$, then if $a_1 < a_2 < \cdots < a_l$, where $l > c_1 n$, there always exists an m such that the number of solutions of

$$m = \sum_{j=1}^{k} a_{i_j}{}^k$$

is greater than c_2. The proof is similar to the proof of (61) but is considerably more tricky.

Turán and I conjectured that if $a_1 < a_2 < \cdots, a_k < ck^2$, is an infinite sequence of integers, then the number of solutions of $n = a_i + a_j$ cannot be bounded. I could only prove that the sums $a_i + a_j$ cannot all be different [43]. One would expect that $a_l < cl^k$ implies that the sums of the a's taken k at a time cannot all be different. Unfortunately the proof works only for even k, and though the result is undoubtedly true for odd k, I cannot prove it. The reason for this difficulty is very simple. For $k = 2$, if the sums $a_i + a_j$ are all distinct, then the differences $a_i - a_j$ are also all distinct. It is easy to see that $a_k < ck^2$ implies that the number of solutions $S(x)$ of $a_i - a_j \leq x$ satisfies $\overline{\lim} S(x)/x = \infty$, and therefore the differences $a_i - a_j$ cannot all be distinct. This argument breaks down for $k = 3$. For further problems and results on this subject I have to refer to my paper on unsolved problems [2a] and to the interesting review article by Stöhr [43].

Varnavides [40] proved using Roth's theorem that if $1 \leq a_1 < \cdots < a_l \leq n$, where $l > \alpha n$, $\alpha > 0$ fixed, n sufficiently large, then the a's contain more than $c_\alpha n^2$ arithmetic progressions of three terms. Except for the value of c_α this result is best possible since the total number of arithmetic progressions $0 < a < a + d < a + 2d \leq n$ is:

$$\binom{\left[\frac{n}{2}\right]}{2} + \binom{\left[\frac{n+1}{2}\right]}{2} = \frac{n^2}{4} + O(n).$$

It would be interesting to determine the best value of c_α and to find the structure of the extremal sequence.

Many problems of combinatorial and numbertheoretical nature are discussed in my paper on unsolved problems [2a], and here I only wish to mention a few of them in which some progress has been made since I wrote the paper.

Denote by $a_1 < \cdots < a_k \leq x$ a sequence of integers for which all the sums $\sum_{i=1}^{k} \epsilon_i a_i$, $\epsilon_i = 0$ or 1, are distinct, and put

$$\max k = A(x).$$

Many years ago I asked whether

$$A(x) = \frac{\log x}{\log 2} + O(1) \tag{64}$$

holds. (64) seems surprisingly resistant to any attack.

I also asked whether $A(2^k) \geq k + 2$ is possible. This was answered affirmatively a few years ago by Guy and Conway (independently). Their example is unpublished. Moser and I proved (see [2], *Colloque . . . Bruxelles*, 136–134) that

$$A(x) < \frac{\log x}{\log 2} + \frac{(1+\epsilon)\log\log x}{2\log 2},$$

and recently Moser showed (to appear in the report of the A.M.S. Pasadena Conference, 1963) that

$$\sum_{i=1}^{k} a_i^2 \geq \frac{4^k - 1}{3}, \tag{65}$$

with equality only for $a_i = 2^{i-1}$. Moser easily deduces from (65) that

$$A(x) < \frac{\log x}{\log 2} + \frac{\log\log x}{2\log 2} + O(1).$$

This is the best upper bound for $A(x)$ known up to the present. It is quite easy to see that if $a_1 < a_2 < \cdots$ is an infinite sequence of integers for which all the sums $\sum_i \epsilon_i a_i$, $\epsilon_i = 0$ or 1, are distinct then for infinitely many i, $a_i \leq 2^{i-1}$. Another somewhat related result states that if $A(x)$ denotes the number of solutions of $\sum_i \epsilon_i a_i \leq x$ and if $A(x) = x + O(1)$ then $a_i = 2^{i-1}$ for $i \geq i_0$. (This is proved in a paper, which will soon appear in *Acta Arithmetica*, by P. Erdös, B. Gordon, L. A. Rubel, and E. Straus.)

Lorenz [44] proved the following conjecture of Straus and myself: Let $a_1 < a_2 < \cdots$ be an infinite sequence of integers; then there always exists a sequence $b_1 < b_2 < \cdots$ of density 0 such that every integer n can be written in the form $a_i + b_j$. In particular he proved that if the a's are the primes then the b's can be chosen so that $B(x) < c(\log x)^3$ $\left(B(x) = \sum_{b_i \leq x} 1\right)$. By using probabilistic arguments [44] I improved this to $B(x) < c(\log x)^2$. The prime number theorem trivially implies $B(x) \geq (1 + o(1))\log x$ and I cannot disprove that $B(x) = (1 + o(1))\log x$. In 1956 Hanani stated the following conjecture: If $a_1 < \cdots$; $b_1 < \cdots$ are two infinite sequences such that every integer can be written in the form

$a_i + b_j$ then

$$\limsup A(x)B(x)/x > 1 \tag{66}$$

Special cases of (66) were proved by Narkiewicz [44]. Recently Danzer disproved (66) (Danzer's paper has just appeared in *J. für reine und angew. Math*). It easily follows from the result of Narkiewicz that to every $\epsilon > 0$ there is an infinite sequence $x_i \to \infty$ such that

$$A(x_i)B(x_i) - x_i > x_i^{1-\epsilon}. \tag{67}$$

The example of Danzer implies the existence of two sequences satisfying

$$x \leq A(x)B(x) \leq x + o(x). \tag{68}$$

There is a gap between (67) and (68) which as far as I know has not yet been filled. Danzer and I conjectured that (68), for two sequences such that every integer can be expressed as $a_i + b_j$, would imply that

$$A(x)B(x) - x \to \infty,$$

but as far as I know this has not yet been proved.

Lorenz's result implies that there exists a sequence $a_1 < a_2 \cdots$ satisfying $A(x) < cx \log\log x/\log x$ such that every integer is of the form $2^k + a_i$. One would expect that this can be improved to $cx/\log x$, but this seems to present unexpected difficulties.

Davenport and I [45] proved that if $a_1 < \cdots$ is an infinite sequence of positive lower density then there exists an infinite subsequence $a_{i_1}, a_{i_2} \cdots$ satisfying $a_{i_k} \mid a_{i_{k+1}}$. I conjectured that there are infinitely many triples a_i, a_j, a_l of distinct integers of the sequence satisfying $[a_i, a_j] = a_l$. This would follow from the following purely combinatorial theorem: Let $A_1, \ldots A_r$ be subsets of a set S of n elements and assume that there are no three distinct sets A_i, A_j, A_l for which

$$A_i \cup A_j = A_l.$$

Put $\max r = f(n)$. Then

$$f(n) = o(2^n). \tag{69}$$

Recently Sárközy and Szemerédi proved (69) (unpublished);

in fact they showed that $f(n) < c2^n/\log\log n$. Perhaps

$$f(n) < c2^n/\sqrt{n},$$

in fact perhaps $f(n) = (1 + o(1))\binom{n}{\left[\frac{n}{2}\right]}$. Thus the above conjecture about triples is now proved. A well-known combinatorial theorem of Sperner [46] should be mentioned here: If the A_i's are such that no one of them contains any other, then their number is at most $\binom{n}{\left[\frac{n}{2}\right]}$. This theorem has many applications in number theory and analysis.

Turán and I conjectured that if $a_1 < \cdots$ is an infinite sequence of integers, and if $F(n)$ denotes the number of solutions of $a_i + a_j \leq n$, then

$$F(n) = cn + O(1)$$

is impossible. Fuchs and I [47] proved that for $c > 0$

$$F(n) = cn + o\left(\frac{n^{1/4}}{(\log n)^{1/2}}\right) \tag{70}$$

is impossible. In the case $a_k = k^2$ Hardy and Landau [47] proved that

$$F(n) = \pi n + o(n \log n)^{1/4} \tag{71}$$

is impossible. This is the classical problem of the number of lattice points in a large circle. It has been conjectured that in (71) the error term is $o(n^{1/4+\epsilon})$, but this seems very deep. It is surprising that in our much more general case we obtain a lower bound for the error term which is nearly as good as (71) and our proof is very much simpler. Recently Jurkat proved that the error term in (70) cannot be $o(n^{1/4})$ (unpublished). Fuchs and I suspected a long time ago that a sequence $a_1 < \cdots$ can be constructed for which

$$F(n) = cn + O(n^{1/4}); \tag{72}$$

this would show that Jurkat's result is best possible, but we have not succeeded in constructing a sequence satisfying (72). Bateman, Kohlbecker, and Tull [47] generalized (70) by replacing c by a slowly oscillating function, but as far as I know the following simple

conjecture has not yet been proved: There does not exist a sequence $a_1 < a_2 \cdots$ for which the number of solutions of $a_i + a_j + a_s \leq x$ is of the form $cx + 0(1)$. My proof with Fuchs breaks down.

Heilbronn and I (our paper will appear in *Acta Arithmetica* [added in proof: 9 (1964), 149–159]) proved that if $a_1, \ldots, a_k$ are distinct residues (mod p), and $k > 3.6^{1/2}\sqrt{n}$, then every residue class (mod p) can be written in the form

$$\sum_{i=1}^{k} \epsilon_i a_i, \qquad \epsilon_i = 0 \text{ or } 1.$$

This result probably holds for $k > 2\sqrt{n}$. To show this it would be sufficient to show that if $a_1, \ldots, a_k$ are k distinct residues (mod p) then the number of distinct residues which can be written as the sum of at most r distinct a's is at least

$$\min\,(p, rk - r^2 + 1). \tag{73}$$

Taking the a's to be the residues $-\left[\frac{k-1}{2}\right], \ldots, +\left[\frac{k}{2}\right]$ we see that (73) if true is best possible. (73) is not even known for $r = 2$. A special case of a well-known theorem of Cauchy-Davenport [48] states that the number of distinct residues which can be written as the sum of r a's (not necessarily distinct) is at least

$$min\,(p, rk - r + 1).$$

Heilbronn and I further proved that if $k/p^{2/3} \to \infty$ then the number of solutions of

$$\sum_{i=1}^{k} \epsilon_i a_i = u(\text{mod } p), \qquad \epsilon_i = 0 \text{ or } 1$$

is $(1 + o(1))2^k/p$. The condition $k/p^{2/3} \to \infty$ is best possible, and $k > cp^{2/3}$ does not suffice.

We further conjectured that if $a_1, \ldots, a_k$ are distinct residues (mod n) and $k > cn^{1/2}$, then

$$\sum_{i=1}^{k} \epsilon_i a_i \equiv 0(\text{mod } n), \qquad \epsilon_i = 0 \text{ or } 1 \tag{74}$$

is always solvable. Perhaps (74) is solvable for every $c > \sqrt{2}$

if $n > n_0(c)$. Flohr and I could only prove that (74) is solvable if

$$k > n^{\gamma+\epsilon}, \qquad \gamma = \frac{1}{1 + \log 2/\log 3};$$

our proof is unpublished.

A famous unsolved problem in number theory asks for the estimation of the least quadratic nonresidue of p. This problem goes back to Gauss who proved that the least quadratic nonresidue is $< 2p^{\frac{1}{2}} + 1$ if $p \equiv 1 (\mathrm{mod}\ 8)$; he used this estimation in his first proof of the law of quadratic reciprocity [49]. The first result which used the modern methods of analytic number theory is due to Vinogradov [49], who proved that the least quadratic nonresidue $n_2(p)$ satisfies

$$n_2(p) < cp^{\frac{1}{2}\sqrt{e}}(\log p)^2. \tag{75}$$

Davenport and I [49] improved the exponent of $\log p$ to $1/\sqrt{e}$. The first significant improvement on (75) was found by Burgess [49] who proved

$$n_2(p) < cp^{\frac{1}{4}\sqrt{e}}(\log p)^{\alpha}. \tag{76}$$

The ingenious proof of Burgess uses the following deep result of A. Weil [50]. Let $f(x)$ be an irreducible polynomial of degree n; then $\left(\left(\frac{f(x)}{p}\right)\right.$ is the Legendre symbol$\left.\right)$

$$\left|\sum_{x=0}^{p-1}\left(\frac{f(x)}{p}\right)\right| < (n-1)\sqrt{p}. \tag{77}$$

André Weil used the methods of algebraic geometry in the proof of (74). For polynomials of degree 3 and 4, (77) was proved by Hasse [50] and weaker inequalities that (77) were proved by Davenport, Mordell, and others [50].

It seems certain that $n_2(p) < p^{\epsilon}$ and in fact perhaps $n_2(p) < c \log p$. Turán observed that $n_2(p) > c' \log p$.

Linnik [49] proved that there is a c_ϵ so that there are at most c_ϵ primes in $x < p < x^2$ which do not satisfy $n_2(p) < p^{\epsilon}$. Linnik developed his famous large sieve for the purpose of proving this result.

Davenport and I [49] observed the trivial result that there exists a constant $c > 0$ so that every interval of length $cp^{\frac{1}{2}}$ contains

both a residue and a nonresidue (mod p). We were not able to prove that this holds for every $c > 0$, but Burgess [49] proved the stronger result that every interval of length $p^{1/4+\epsilon}$ contains both a residue and a nonresidue. It seems probable that $p^{1/4+\epsilon}$ can be replaced by p^{ϵ} or even by $c \log p$.

I proved [51] that

$$\sum_{p<x} n_2(p) = (1 + o(1)) \frac{x}{\log x} \sum_{k=1}^{\infty} \frac{p_k}{2^k}. \tag{78}$$

Very likely, if $n_k(p)$ denotes the least kth power nonresidue of p, one has

$$\sum_{p<x} n_k(p) = c_k \frac{x}{\log x} + o\left(\frac{x}{\log x}\right), \tag{79}$$

but my knowledge of algebraic number theory was not sufficient to enable me to prove (79). The proof of (78) is surprisingly complicated; I needed to use the prime number theorem for airthmetic progressions, Brun's method, and the large sieve of Linnik and Rényi.

Denote by $f(\epsilon, p)$ the smallest integer such that, for every $l \geq f(\epsilon, p)$

$$\sum_{n=1}^{l} \left(\frac{n}{p}\right) < \epsilon l.$$

I expect that

$$\sum_{p<x} f(\epsilon, p) = (1 + o(1)) c_{\epsilon} \frac{x}{\log x}, \tag{80}$$

but I do not see how to attack (80).

Denote by $r(p)$ the least primitive root of p. Vinogradov proved $r(p) < p^{\frac{1}{2}+\epsilon}$; Hua, Shapiro, and I [52] improved this to

$$(r(p-1))^{c} p^{\frac{1}{2}}.$$

The first significant improvement on Vinogradoff's result is due to Burgess [52] who proved

$$r(p) < p^{\frac{1}{4}+\epsilon}.$$

Very likely $r(p) < c \log p$. Ankeny [52] deduced from the general-

ized Riemann Hypothesis that $n_2(p) < c(\log p)^2$, and obtained a weaker result for $r(p)$.

It seems very hard to prove that

$$\sum_{p<x} r(p) = (1 + o(1)) \frac{cx}{\log x},$$

in fact I cannot even show that $\lim r(p) < \infty$. Artin conjectured that there are infinitely many primes p for which 2 is a primitive root, in fact he made a plausible conjecture about their density. As far as I know it is not even known whether to every p there is a prime $q < p$ which is a primitive root of p.

In the last 35 years significant new results were obtained in the additive theory of prime numbers, also in Waring's problem, but I do not wish to speak much of these since several excellent books discussed them recently in great detail. I only wish to state the closest approaches known to the famous Goldbach conjecture: Selberg and Wang [53] proved, using Selberg's improvement of Brun's method, that every sufficiently large integer can be written in the form $a + b$ where a has at most 2 prime factors and b at most 3. Rényi proved [53] that there is an absolute constant c such that every integer is the sum of a prime and an integer which has at most c prime factors. Rényi used the large sieve of Linnik and Rényi and new results about the distribution of the roots of L-functions. I have been informed that recently Barban proved that every large integer can be written in the form $p + a$ where $\nu(a) \leq 4$. One of the main new ideas of Barban's proof is a remarkable improvement of the Linnik-Rényi large sieve.

Another remarkable recent result is due to Linnik [31]. He proved by his dispersion method the following conjecture of Hardy and Littlewood: The number of solutions of $n = p + u^2 + v^2$ is of the form

$$(1 + o(1))c_n \frac{n}{\log n}, \qquad c_n > \frac{c}{\log \log n}$$

where c_n is a complicated constant which depends only on n. He also obtains asymptotic formuls for sums of the form $\left(d_k(m) = \sum_{x_1 \cdots x_k = m} 1\right)$

$$\sum_{m \leq n} d_{k_1}(m)\, d_{k_2}(m + a)$$

and

$$\sum_{m \leq n} d_{k_1}(m)\, d_{k_2}(n-m)$$

for $k_1 = 2$ and k_2 arbitrary. For the details and the history of this problem I have to refer to Linnik's book.

Many of the results mentioned in this chapter can be proved by Brun's method, which is perhaps our most powerful elementary tool in number theory. Recently Selberg [12] obtained a significant improvement of Brun's method and in a certain sense showed that further improvement is impossible beyond a certain limit. As far as I know the following question has not yet been investigated. Determine or estimate the smallest $f_1(x)$ with the following property: there exists a set of residue classes $a_i \pmod{p_i}$, for $p_i < f_1(x)$, such that every $1 \leq u \leq x$ satisfies at least one of the congruences $u \equiv a_i \pmod{p_i}$. Similarly for $f_2(x)$, defined as follows: there is a set of residue classes $a_i \pmod{p_i}$, $p_i < f_2(x)$ so that the number of integers $u \leq x$ which do not satisfy any of the congruences $u \equiv a_i \pmod{p_i}$ is $o(x/\log x)$.

Here I would like to call attention to another problem on sieve methods which as far as I know has not yet been investigated. The essential result proved by Viggo Brun was the following. Let $p < n^{\epsilon}$, $\epsilon = \epsilon(k)$ and consider $k_p \leq k$ congruences:

$$x \equiv a_j^{(p)} \pmod{p}, \qquad 1 \leq j \leq k_p \leq k. \tag{81}$$

Then the number of integers $x \leq n$ which do not satisfy any of the congruences (81) is between

$$c_1 n \prod_{p<n^{\epsilon}} \left(1 - \frac{k_p}{p}\right) \quad \text{and} \quad c_2 n \prod_{p<n^{\epsilon}} \left(1 - \frac{k_p}{p}\right). \tag{82}$$

(82) was improved by Selberg [12] in two ways; he permitted a larger choice of $\epsilon = \epsilon(k)$, and he brought c_1 and c_2 closer together.

Linnik and Rényi investigated the other extreme; in their case the number of congruences (81) is very large, and roughly speaking they prove that if "many" integers are given up to x then, with the exception of a few primes, each residue class mod p contains "nearly" the same number of integers if we neglect a "few" exceptional residue classes. For a precise statement I have to refer to the papers of Rényi and Linnik [53].

As far as I know nobody investigated what happens if in (81) the

number of congruences increases with p but not very quickly, say like $\log p$ or like $\log \log p$. I have not been able to find a reasonable application for the estimation which would correspond to (82) and this may be the reason why this question was neglected.

Now I would like to call attention to another group of problems on congruences. A set of congruences:

$$(83) \qquad a_i \pmod{n_i}, \qquad n_1 < n_2 < \cdots < n_k$$

is called a covering set if every integer satisfies at least one of the congruences (83). The simplest covering set is $0 \pmod 2$, $0 \pmod 3$, $1 \pmod 4$, $5 \pmod 6$, $7 \pmod{12}$. I asked if for every choice of n_1 there is a covering set (83). This problem seems very difficult and has been solved only for $n_1 \leq 8$ (by Selfridge and others). Many similar questions can be asked, for example, it is not known if there exists a covering set with all the n_i odd. A simple but not quite trivial result about covering sets of congruences states that $\sum_{i=1}^{k} \frac{1}{n_i} > 1$ [54].

Two congruences are called disjoint if no integer satisfies both of them. Stein and I asked: let (83) be a system of pairwise disjoint congruences for which $n_k \leq x$. Put $\max k = f(x)$. We conjectured that $f(x) = o(x)$. We proved that for every $\epsilon > 0$ and $x > x_0(\epsilon)$

$$f(x) > x \exp\left(-(\log x)^{\frac{1}{2}-\epsilon}\right).$$

It is surprising that the proof of $f(x) = o(x)$ presents difficulties and this perhaps due to our overlooking an obvious idea [54].

Stein conjectured that if (83) is a disjoint system then there always exists a $0 < u \leq 2^k$ for which $u \not\equiv a_i \pmod{n_i}$, $(i = 1, 2, \ldots, k)$. This conjecture was recently proved by Selfridge (unpublished). The system

$$2^{i-1} \pmod{2^i}, \qquad 1 \leq i \leq k$$

shows that this conjecture is best possible.

I conjectured that if $a_i \pmod{n_i}$, $1 \leq i \leq k$ is any system of congruences such that there is a u for which $u \not\equiv a_i \pmod{n_i}$ $(i = 1, \ldots, k)$, then there is such a u for which $0 < u \leq 2^k$. I could only prove that there is such a u for which $0 < u < f(k)$ where $f(k)$

depends only on k, but I did no give an explicit estimation for $f(k)$ [54].

Recently Elliott (in publication in *Quart. J. of Math.*) proved that if $a_1 < a_2 < \cdots < a_k < n$ and $k > cn/(\log n)$, where $c > 2$, then there exists a prime q such that every residue class (mod q) is represented among the a_i. He uses Selberg's sieve.

8. I now refer briefly to a body of recent work on Diophantine equations and Diophantine inequalities in many variables. We consider equations first.

The treatment of equations of additive type, such as

$$f(x_1) + \cdots f(x_n) = N$$

where $f(x)$ is an integer-valued polynomial, is possible by the Hardy-Littlewood method, and presents no essentially new difficulty. In particular a homogeneous additive equation:

$$a_1 {x_1}^k + \cdots + a_n {x_n}^k = 0$$

always has an infinity of solutions, provided n is greater than a suitable function of k, and provided $a_1, \ldots, a_n$ are not all of the same sign if k is even.

The general homogeneous equation

$$f(x_1, \ldots, x_n) = 0$$

of degree k, offers much more difficulty. The first method of reducing such an equation to additive equations (in a smaller number of variables) was given by Richard Brauer in 1945. This method was not directly applicable in the rational number field, because it required the solubility of all additive equations of every degree $k' \leq k$, and this cannot be ensured for even values of k'. But Lewis (1957) modified Brauer's method to obtain a proof that when $k = 3$ the equation is always soluble if n is sufficiently large. Birch (1957) generalized Brauer's method to prove the following remarkable theorem: every system of simultaneous equations, of odd degrees $k_1, \ldots, k_r$, is soluble in integers (not all 0) provided n is greater than a certain function of $k_1, \ldots, k_r$.

Davenport (1959 and 1963) attacked the problem of a single homogeneous cubic equation directly by the Hardy-Littlewood method, and proved that the condition $n \geq 16$ is sufficient to ensure solubility. It is conjectured that $n \geq 10$ suffices and it is known

that this would be best possible. A treatment on similar general lines of homogeneous equations of higher degree, and simultaneous systems of such equations, was given by Birch (1962), but here it becomes necessary (and it is indeed essential) to impose further conditions.

A connected account of much of this work, with references, is available in Davenport's notes: *Analytic Methods for Diophantine Equations and Diophantine Inequalities* (Ann Arbor Publishers, 1963).

There is a close connection between these problems and the problem of the solubility of homogeneous equations in p-adic numbers. In several recent papers, Birch and Lewis have established, for particular values of k, the conjecture of Artin that for such equations the condition $n > k^2$ is sufficient to ensure solubility.

As regard Diophantine inequalities, the principal result of a general character proved so far is the following: if $Q(x_1, \ldots, x_n)$ is any indefinite quadratic form with real coefficients, then the inequality

$$|Q(x_1, \ldots, x_n)| < \epsilon$$

is soluble for every $\epsilon > 0$ provided $n \geq 21$. (For references see Davenport's notes, mentioned above.) The proof is complicated. It is conjectured that $n \geq 5$ suffices, and it may even be true that $n \geq 3$ would suffice if one excluded forms which are proportional to forms with integral coefficients. A similar result (but with a very large lower bound for n) can probably be proved for cubic forms, but further extension seems to present great difficulties.

REFERENCES

1. I. M. Vinogradov, *The method of trigonometrical sums in the theory of numbers*, translated from the Russian by K. F. Roth and Ann Davenport, New York (1954) Interscience Publishers.
L. K. Hua, *Additive Primzahltheorie*, Leipzig (1959) Teubner. For the older results see the excellent book of E. Landau, *Vorlesungen über Zahlentheorie*, vols. 1–3; Leipzig (1927).
Hardy and Wright, *Introduction to the theory of numbers*, Oxford, Clarendon Press.
L. K. Hua, Die Abschätzung von Exponentialsummen und ihre Anwendung in der Zahlentheorie, *Enz. Math. Wiss.*, Band I 2 Heft 13 Teil I (1959).
For the most modern results on prime number theory, see K. Prachar, *Primzahlverteilung*, Berlin (1957) Springer.

2. M. Kac, Probability methods in some problems of analysis and number theory, *Bull. Amer. Math. Soc.*, **55** (1949) 641–665.
Kubilius, Probabilistic methods in the theory of numbers (Russian), *Uspehi Matem. Nauk*, **11** (1956) 31–66. See also Probabilistic methods in the theory of numbers, *Transl. Math. Monog.* **11** (1964) Amer. Math. Soc.
P. Erdös, On additive arithmetical functions and applications of probability to number theory, *Proc. Int. Congress of Math. Amsterdam*, **3** (1954) 13–19.
P. Erdös, On the distribution function of additive arithmetical functions and on some related problems, *Rend. Sem. Math. Fis. Milano*, **27** (1958) 3–7.
E. V. Novoselov, A new method in probabilistic number theory, *Izv. Akad. Nauk U.S.S.R.*, **28** (1964) 307–364.
See also a forthcoming review article by me in the *London Math. Soc. Journal*, On applications of probability methods to number theory. For applications of probability methods to additive number theory see, for example, A. Rényi, Probabilistic methods in number theory, *Proc. I.C.M. Edinburgh* (1958) 529–539.
P. Erdös, Problems and results in additive number theory, *Colloque sur la théorie des nombres*, 127–137, Bruxelles (1955); P. Erdös and A. Rényi, Additive properties of random sequences of positive integers, *Acta Arithmetica*, **6** (1960) 83–110; see also a forthcoming book on sequences of integers by K. F. Roth and H. Halberstam.

2a. P. Erdös, On unsolved problems, *Publ. Math. Inst. Hung. Acad.*, **6** (1961) 221–254.

3. A. Selberg, An elementary proof of the prime number theorem, *Annals of Math.*, **50** (1949) 305–313.
P. Erdös, On a new method in elementary number theory which leads to an elementary proof of the prime number theorem, *Proc. Nat. Acad. Sci. U.S.A.*, **35** (1949) 379–384.
P. Erdös, On a Tauberian theorem connected with the new proof of the prime number theorem, *Journ. Ind. Math. Soc.*, **13** (1949) 133–147.
H. N. Shapiro, Tauberian theorems and elementary prime number theory, *Comm. Pure Appl. Math.*, **12** (1959) 579–610.

4. A. Beurling, Analyse asymptotique de la distribution des nombres premiers generalisés, *Acta Math.*, **68** (1937) 255–299.
B. Nyman, A general prime number theorem, *ibid.* **81** (1949), 299–307.
P. Malliavin, Sur la reste de la loi asympotique de répartition des nombres premiers généralisés de Beurling, *ibid.* **106** (1961) 281–298.

5. E. Bombieri, Sulle formule di A. Selberg generalizzate per classi di funzioni aritmetiche e le applicazioni al problema del resto nel Primzahlsatz, *Riv Mat. Univ. Parma Ser II*, **3** (1962), 393–440. This paper has a very good bibliography of the literature of the elementary proofs of the prime number theorem and their generalizations.
E. Wirsing, Elementare Beweise des Primzahlsatzes mit Restglied I, *J. reine und angew. Math.*, **211** (1962), 205–214; and **214–215** (1964), 1–18.

6. T. V. Nevanlinna, Über den Elementaren Beweis des Primzahlsatzes, *Soc. Sci. Fenn. Comment. Phys. Math. 27*, **3** (1962); and Über die elementaren

Beweise der Primzahlsätze und deren aequivalente Fassungen, *Ann. Acad. Sci. Fennicae Ser. A. I. Math*, **343** (1964).

7. A. E. Ingham, Some Tauberian theorems connected with the prime number theorem, *J. London Math. Soc.*, **20** (1945) 171–180.
8. N. M. Korobov, Weyl's sums estimates and the distribution of primes, *Doklady Akad. Nauk. U.S.S.R.*, **123** (1959) 28–31; I. M. Vinogradov, *Akad. Nauk. U.S.S.R. Ser Mat.*, **22** (1958) 161–164.
9. W. Haneke, Verschärfung der Abschätzung von $\xi(\frac{1}{2} + it)$, *Acta Arith.* **8** (1962/63), 357-430.
10. H. Cramér, On the order of magnitude of the difference between consecutive prime numbers, *Acta Arith.*, **2** (1936) 23–46.
11. P. Erdös, On the difference of consecutive primes, *Quart. J. Math. Oxford*, **6**, 124–128.
    T. H. Chang, Über aufeinanderfolgende Zahlen von denen jede mindestens einer von n linearen Kongruenzen genügt, deren Moduln die ersten n Primzahlen sind, *Schriften Math. Sem. und Inst. Angew. Math. Univ. Berlin*, **4** (1938) 35–55.
    R. A. Rankin, The difference between consecutive prime numbers, *J. London Math. Soc.*, **13** (1938) 242–247.
    A. Schönhage, Eine Bemerkung zur Konstruktion grosser Primzahllücken, *Arch. Math.*, **14** (1963) 29–30.
    R. A. Rankin, The difference between consecutive prime numbers V, *Proc. Edinburgh Math. Soc.*, **13** (1963) 331–332.
12. G. H. Hardy and J. E. Littlewood, Some problems of partitio numerorum III, *Acta Math.*, **44** (1923) 1–70.
    A. Selberg, On elementary methods in prime number theory and their limitations, *Den 11-te Skand. Mat. Kongress* (1952) 13–22; see also The general sieve method and its place in prime number theory, *Proc. Int. Congr. Math. Harvard*, **1** (1950) 286–292.
13. G. Ricci, Recherches sur l'allure de la suite $(p_{n+1} - p_n)/\log p_n$, *Colloque sur la théorie des nombres, Bruxelles* (1955) 93–96; Sull, andamento della differenza di numeri primi consecutivi, *Riv. Math. Univ. Parma*, **5** (1954) 3–54.
14. P. Erdös, The difference of consecutive prime numbers, *Duke Math. J.*, **6** (1940) 438–441.
    R. A. Rankin, The difference between consecutive prime numbers, II, III, IV, *Proc. Cambridge Phil. Soc.*, **36** (1940) 255–266; *J. London Math. Soc.*, **22** (1947) 226–230; *Proc. Amer. Math. Soc.*, **1** (1950) 143–150.
15. P. Erdös and P. Turán, On some new questions in the distribution of prime numbers, *Bull. Amer. Math. Soc.*, **54** (1948) 271–278.
    P. Erdös and A. Rényi, Some problems and results on consecutive primes, *Simon Stevin* (1949) 115–126.
    K. Blarner, Zur Abschätzung von Reihen deren Glieder von rationalen Funktionen einer festen Anzahl sukzessiver Primzahlen gebildet werden, *Monatshefe für Math.*, **68** (1964) 1–16.
    W. Schwarz, Weitere mit einer Methode von Erdös-Prachar erzielte Ergebnisse, *ibid.* 75–80.

16. W. Sierpinski, Remarque sur la répartition des nombres premiers, *Colloquium Math.*, **1** (1948) 193–194.
A. Walfisz, Isolierte Primzahlen, *Doklady Akad. Nauk. U.S.S.R.*, **90** (1953) 711–713.
K. Prachar, Über ein Resultat von A. Walfisz, *Monatshefte Math.*, **58** (1954) 114–116.
P. Erdös, On some applications of Brun's method, *Acta Szeged*, **13** (1949) 57–63.
17. P. Erdös and K. Prachar, Sätze und Probleme über p_k/k, *tbh. Math. Sem. Univ. Hamburg*, **25** (1962) 51–56.
18. P. Erdös, Problems and results on the differences of consecutive primes, *Publ. Math. Debrecen* **1** (1949) 33–37.
19. P. Erdös, On the difference of consecutive primes, *Bull. Amer. Math. Soc.*, **54** (1948) 885–889.
20. P. Erdös, Some problems and results in elementary number theory, *Publ. Math. Debrecen*, **2** (1951) 103–109.
H. E. Richert, On the difference between consecutive squarefree numbers, *J. London Math. Soc.*, **29** (1954) 16–20. [*Added in proof:* It was pointed out to me that Richert's result was improved by Rankin (*Quart. J. Math., Ser.* 2, **6** (1955) 147–153) who replaced the exponent $\frac{2}{3}$ by 0.22198215. The best result to date is due to P. G. Schmidt who in his dissertation (Göttingen, 1964) obtained the exponent $\frac{1095\,56}{4944\,19} = 0.22158534$.]
C. Hooley, On the difference of consecutive numbers prime to n, *Acta Arith.*, **8** (1963) 343–347.
21. P. Erdös, On the integers relatively prime to n and on a number theoretic function considered by Jacobsthal, *Math. Scand.*, **10** (1962) 163–170.
22. A. Selberg, An elementary proof of Dirichlet's theorem about primes in an arithmetic progression, *Ann. of Math.*, **50** (1949) 297–307.
A. Selberg, An elementary proof of the prime number theorem for arithmetic progressions, *Canadian J. Math.*, **2** (1950) 66–78.
H. N. Shapiro, On primes in arithmetic progressions I, *Ann. of Math.* (2) (52) (1950) 217–230; II ibid. (2) (52) (1950) 231–243.
W. E. Briggs, An elementary proof of a theorem about the representation of primes by quadratic forms, *Canadian J. Math.* **6** (1954) 353–363.
H. N. Shapiro, An elementary proof of the prime ideal theorem, *Comm. Pure Appl. Math.*, **2** (1949) 309–323.
W. Forman and H. N. Shapiro, Abstract prime number theorem, *ibid.* **7** (1954) 587–619.
23. For an easily accessible exposition of the work of Linnik and Rodosskij, see K. Prachar's book *Primzahlverteilung*, Berlin (1957) Springer.
P. Turán, On a density theorem of Ju. V. Linnik, *Publ. Math. Inst. Hung. Acad.*, **6** (1961) 165–178.
S. Knapowski, On Linnik's theorem concerning exceptional L-zeros, *Publ. Math. Debrecen*, **9** (1962) 168–178.
24. P. Turán, Über die Primzahlen der arithmetischen Progression, *Acta Szeged*, **8** (1937) 226–235.
P. Erdös, On some applications of Brun's method, *ibid.* **13** (1949) 57–63.

I. M. Vinogradov, Über die Abschätzung trigonometrischer Summen mit Primzahlen, *Izv. Akad. Nauk, U.S.S.R. Ser. Mat*, **12** (1948) 225–248.

25. S. Knapowski and P. Turán, Comparative prime-number theory I–VIII, *Acta Math. Hung.* **XIII** (1962), **XIV** (1963) and also forthcoming papers in *Acta Arithmetica*. A first exposition of the methods in question can be found in Turán's book *Eine neue Methode in der Analysis und deren Anwendungen;* a completely rewritten English edition will be published in the Interscience Tracts series. For the older literature, see the papers of Turán and Knapowski and the book of Turán.
26. S. Knapowski, On sign-changes in the remainder-term in the prime number formula, *J. London Math. Soc.*, **36** (1961) 451–460.
27. F. V. Atkinson, On sums of powers of complex numbers, *Acta Hung. Acad.*, **12** (1961) 185–188.
28. W. E. Briggs, Prime-like sequences generated by a sieve process, *Duke Math. J.*, **30** (1963) 297–311. This paper contains the sharpest results known up to the present and has references to the previous literature.
29. D. Hawkins, The random sieve, *Math. Mag.*, **31** (1957–1958) 1–3.
30. P. Erdös, On a problem of S. Golomb, *J. Australian Math. Soc.*, **2** (1961) 1–8.
31. P. Turán, Uber einige Verallgemeinungen eines Satzes von Hardy und Ramanujan, *J. London Mat. Soc.*, **11** (1936) 125–133.
    S. Ramanujan, *Collected papers*, 262–275.
    H. Halberstam, On the distribution of additive number-theoretic function, II and III, *J. London Math. Soc.*, **31** (1956) 1–11, 14–31.
    P. Erdös, On the normal number of prime factors of $p - 1$ and some related problems concerning Euler's function, *Quart. J. Math. Oxford*, **6** (1935) 205–213.
    E. C. Titchmarsh, On a divisor problem, *Rend. Circ. Mat. Palermo*, **54** (1930) 414–419.
    C. B. Haselgrove, Some theorems in the analytic theory of numbers, *J. London Math. Soc.*, **26** (1951) 273–277.
    Yu. V. Linnik, The dispersion method in binary additive problems, *Trans. Math. Monog.* **4** (1963) Amer. Math. Soc.
32. J. G. van der Corput, Une inégalité rélative au nombre des diviseurs, *Proc. Neder. Akad. Wet. Amsterdam*, **42** (1939) 547–553.
    P. Erdös, On the sum $\sum_{k=1}^{n} d(f(k))$, *J. London Math. Soc.*, **27** (1952) 7–15.
    R. Bellman, Ramanujan sums and the average value of arithmetic functions, *Duke Math. J.*, **17** (1950) 159–168.
    C. Hooley, On the number of divisors of quadratic polynomials, *Acta Math.*, **110** (1963) 97–114.
33. P. Erdös, On the greatest prime factor of $\prod_{k=1}^{x} f(k)$, *J. London Math. Soc.*, **27** (1952) 379–384. The references to the older literature can be found in this paper.
    P. Erdös and R. Rado, Intersections theorems for systems of sets, *J. London Math. Soc.*, **35** (1960) 85–90.

34. A. Schinzel, On primitive prime factors of $a^n - b^n$, *Proc. Cambridge Phil. Soc.*, **58** (1962) 555–562.
G. Pólya, Zur arithmetischen Untersuchung der Polynome, *Math. Zeitschrift*, **1** (1918) 143–148.
C. Siegel, Über Näheringswerte algebrascher Zahlen, *Math. Annalen*, **84** (1921) 80–99; see also K. Mahler, *Math. Annalen*, **107** (1933) 691–730.
S. Chowla, The greatest prime factor of $x^2 + 1$, *J. London Math. Soc.*, **10** (1935) 117–120.
K. Mahler, Über den grössten Primteiler der Polynome $x^2 \pm 1$, *Arch. Math. og Naturvid.*, **41** (1935) 1–8.
35. I. I. Pjatezkij-Schapiro, On the distribution of primes in the sequences of the form $[f(n)]$, *Mat. Sbornik*, **33** (1953) 559–566.
36. H. Heilbronn, Über die Verteilung der Primzahlen in Polynomen, *Math. Ann.*, **104** (1931) 794–799.
37. P. Erdös, Arithmetical properties of polynomials, *J. London Math. Soc.*, **28** (1953) 416–425.
38. B. L. van der Waerden, Beweis einer Baudetschen Vermutung, *Nieuw Archiefvoor Wiskunde*, (2) **15** (1928), 212–216.
P. Erdös and R. Rado, Combinatorial theorems on classifications of subsets of a given set, *Proc. London Math. Soc.*, **2** (1952) 438–439.
W. Schmidt, Two combinatorial theorems on arithmetic progressions, *Duke Math. J.*, **29** (1962) 129–140.
P. Erdös, On some combinatorial problems connected with the theorems of Ramsey and Van der Waerden (in Hungarian), *Mat. Lapok*, **14** (1963) 29–37.
R. Rado, Studien zur Kombinatorik, *Math. Zeitsch.*, **36** (1933) 424–480; see also Khintchine's delightful book *Three pearls of number theory*, English translation (Graylock) (1952).
N. G. Tchudakoff, Theory of the characters of number semigroups, Report Internat. Coll on Zeta Function, Bombay (1956), also *J. Indian Math. Soc.* **20** (1956) 11–17.
39. A. Brauer, Uber sequenzen von Potenzresten I and II, *Sitzungsber. Preuss. Akad. Wiss. Phys.-Math. kl.* **19** (1931) 329–341.
40. P. Erdös and P. Turán, On some sequences of integers, *J. London Math. Soc.*, **11** (1936) 261–264.
F. Behrend, On sequences of integers containing no arithmetic progression, *Cas. Mat. fys.*, **67** (1938) 235–238.
R. Salem and D. C. Spencer, On sets of integers which contain no three terms in an arithmetic progression, *Proc. Nat. Acad. Sci. U.S.A.*, **28** (1942) 561–563.
F. Behrend, On sets of integers which contain no three terms in an arithmetical progression, *ibid.* **32** (1946) 331–332.
R. A. Rankin, Sets of integers containing not more than a given number of terms in arithmetical progression, *Proc. Royal Soc. Edinburgh A*, **65** (1960–1961) 332–344.
K. F. Roth, On certain sets of integers, *J. London Math. Soc.*, **28** (1953) 104–109.
P. Varnavides, On certain sets of positive density, *J. London Math. Soc.*, **341** (1959) 358–360.

L. Chowla, There exists an infinity of 3-combinations of primes in A, *P. Proc. Lahore Philos. Soc.* **6**, No. 2 (1944) 15–16.
See also J. G. van der Corput, Über Summen von Primzahlen und Primzahlquadraten, *Math. Ann.* **113** (1939) 1–50.

41. P. Erdös, On the sum and difference of squares of primes, I and II, *J. London Math. Soc.*, **12** (1937) 133–136, 168–171.
42. P. Erdös, On the representation of an integer as the sum of k kth powers, *J. London Math. Soc.*, **11** (1936) 133–136.
S. Chowla and S. S. Pillai, The number of representations of a number as a sum of non-negative nth powers, *Quart. J. Math. Oxford*, **7** (1936) 56–59.
K. Mahler, Note on hypothesis K of Hardy and Littlewood, *J. London Math. Soc.*, **11** (1936) 136–138.
43. A. Stöhr, Gelöste und ungelöste Fragen über Basen der natürlichen Zahlenreihe, I and II, *J. reine u. angew. Math.*, **194** (1955) 40–65, 111–140. Many interesting unsolved problems can be found in this paper. For a review of additive number theory, see H. H. Ostmann, *Additive Zahlentheorie, Ergeb. der Math.*, Heft 7 and 11.
44. G. G. Lorenz, On a problem of additive number theory, *Proc. Am. Math. Soc.*, **5** (1954) 838–841.
P. Erdös, Some results on additive number theory, *ibid.* 847–853.
W. Narkiewicz, Remarks on a conjecture of Hanani in number theory, *Coll. Math.*, **7** (1960) 161–165.
45. P. Erdös and H. Davenport, On sequences of positive integers, *Acta Arithmetica*, **2** (1936) 147–151; and *J. Indian Math. Soc.*, **15** (1951) 19–24.
46. A. Sperner, Ein Satz über Untermengen einer endlichen Menge, *Math. Zeitsch.*, **27** (1928) 544–548.
47. P. Erdös and W. H. J. Fuchs, On a problem of additive number theory, *J. London Math. Soc.*, **31** (1956) 67–73.
P. T. Bateman, E. E. Kohlbeiker, and J. P. Tull, In a theorem of Erdös and Fuchs in additive number theory, *Proc. Amer. Math. Soc.* **14** (1963) 278–284.
48. H. Davenport, On the addition of residue classes, *J. London Math. Soc.*, **10** (1935) 30–32; A historical note, *ibid.* **22** (1947) 100–101.
49. C. F. Gauss, *Disquisitiones arithmeticae*, art. 125 and 129; see also Dirichlet-Dedekind, *Vorlesungen über Zahlentheorie*, 4th edition (1894), p. 116.
A. Brauer, Über den kleinsten quadratischen Nichtrest, *Math. Zeitsch.*, **33** (1931) 162–176.
J. M. Vinogradoff, On the bound of the least non-residue of kth powers, *Trans. Amer. Math. Soc.*, **29** (1927) 218–226; see also *Journal of the Physico-Mathematical Soc. of Perm.* (1919).
H. Davenport and P. Erdös, The distribution of quadratic and higher residues, *Publ. Math. Debrecen*, **2** (1952) 252–265.
D. A. Burgess, The distribution of quadratic residues and nonresidues, *Mathematika*, **4** (1957) 106–112.
Yu. V. Linnik, A remark on the least quadratic non-residue, *Doklady Acad. Sci. U.S.S.R.* (*N.S.*), **36** (1942) 119–120.
50. A. Weil, *Sur les courbes algébriques et les variétés qui s'en déduisent, Actualités Math. Sci. No.* 1041, Paris 1945.

H. Hasse, Theorie der relativ-zyklischen algebraischen Funktionenkörper inbesdondere bei endlichem Konstantenkörper, *J. reine u. angew. Math.*, **172** (1935) 37–59.

H. Davenport, On character sums in finite fields, *Acta Math.*, **71** (1939) 99–121.

51. P. Erdös, Remarks on number theory I, *Mat. Lapok*, **12** (1961) 10–17.
52. P. Erdös and H. N. Shapiro, On the least primitive root of a prime, *Pacific J. of Math.*, **7** (1957) 861–865.

N. C. Ankeny, The least quadratic non-residue, *Annals of Math.*, **55** (1952) 65–71.

D. A. Burgess, On character sums and primitive roots, *Proc. London Math. Soc.*, (3) **12** (1962) 179–192.

Wang Yuan, A note on the least primitive root of a prime, *Science Record, New Ser.*, **3** (1959) 174–207.

53. Wang Yuan, On sieve methods and some of their applications, *Scientia Sinica*, **8** (1959) 357–381.

A. Rényi, On the large sieve of Yu. V. Linnik, *Comp. Mat.*, **8** (1950) 68–75; see also On the probabilistic generalisation of the large sieve of Linnik, *Publ. Math. Inst. Hung. Acad.*, **3** (1958) 199–206.

54. P. Erdös, On a problem of systems of congruences (in Hungarian), *Mat. Lapok*, **3** (1952) 122–128.

S. K. Stein, Unions of arithmetic sequences, *Math. Annalen*, **134** (1957) 289–294.

P. Erdös, External problems in number theory (in Hungarian), *Mat. Lapok*, **13** (1962) 228–255; see also my forthcoming paper in the report of the A.M.S. Pasadena conference (1963). For problems and conjectures on prime numbers, see A. Schinzel and W. Sierpinski, Sur certaines hypothèses concernant les nombres premiers, *Acta Arithmetica*, **4** (1958) 185–207. Several unsolved problems are stated in a recent paper of Sierpinski, On some unsolved problems of arithmetics, *Scripta Math.*, **25** (1960) 125–136; see also P. Erdös, *Quelques problèmes de la théorie des nombres, Monographies de l' Enseignement Math.*, No. 6 (1963), 81–135.

5

On Stochastic Processes

Michel Loève

I. TRADITIONAL SETUP

1. *Introduction.* A characteristic proposition of probability theory, and one of the most beautiful, may be stated in intuitive terms as follows: a gambler plays related or unrelated games of any kind. Except for a miracle, his chances of ruin evaluated in terms of past outcomes approach 100 per cent or 0 per cent according as he will or will not be ruined eventually.

In probability terms, this P. Lévy zero-one law (the first martingales convergence theorem) is as follows: if A is a property of (that is, event defined in terms of) a sequence $X_1, X_2, \ldots$ of random variables, then the conditional probability $P(A|X_1, \ldots, X_n)$ of A given the first n terms of the sequence converges almost surely to the indicator of A ($I_A = 1$ on A and 0 on its complement A^c). Unless otherwise stated, convergence will be for $n \to \infty$.

In measure-theoretic terms, this proposition becomes: let X_1, $X_2, \ldots$ be a sequence of measurable functions on a measure space $(\Omega, \mathcal{A}, P)$ with $P\Omega = 1$. Let $\mathcal{A}_n$ and $\mathcal{A}_\infty$ be the smallest σ-fields for which the finite sequence $X_1, \ldots, X_n$ and the whole sequence $X_1, X_2, \ldots$ are respectively measurable. If $A \in \mathcal{A}_\infty$ and $P(A \mid \mathcal{A}_n)$

is a Radon-Nikodym derivative of the restriction of $P(A \cap \cdot)$ to $\mathcal{A}_n$ with respect to the restriction $P_{\mathcal{A}_n}$ of P to $\mathcal{A}_n$, then $P(A \mid \mathcal{A}_n)$ converges P-almost everywhere to I_A.

The above situation occurs regularly in probability theory: the phenomenological content of a concept or a proposition is intuitively clear and may even sound trivial. The mathematical formulation is quite abstract. The proof may require intricate manipulation of deep mathematical ideas and results, seemingly unmotivated by the intuitive content. However, probability theory developed and continues to develop its own "probabilistic intuition," reflected in its terminology and, especially, in its problems.

The language of probability theory frequently sounds unmathematical and is easily translatable in terms belonging to other branches of mathematics. The origin of its concepts and problems (ancient and new) as well as its immediate applications are frequently extremely concrete. Perhaps because of this and despite its growing importance for other branches of mathematics, most mathematics students receive no foundation in probability theory and many mathematicians are not familiar with its methods and main results and, sometimes, not even with its terminology.

Since this talk is addressed to a general mathematical audience, I shall begin with a short probabilistic vocabulary, as formalized by Kolmogorov (1933), and with the main stochastic structures and three results. This will provide us with a basis for discussion of fields and directions of research, some active, some dormant, and some possible. Probability theory is in full flood: the torrential flow creates new problems, requires new methods, uncovers new phenomena, and deposits new results—even in the most classical cases. Within the limited time at my disposal, it is useless even to try to give a well-rounded picture. What follows will be centered about a few ideas selected with obvious personal bias.

2. *Basic Vocabulary and Notation.* *A probability space* $(\Omega, \mathcal{A}, P)$ is a normalized measure space: $\mathcal{A}$ is a Boolean σ-algebra or σ-field of *events* and the *probability* P is a measure on $\mathcal{A}$ with $P\Omega = 1$. A relation holds almost surely (a.s.) if it holds outside a P-null event.

A *random variable* X is a measurable function on the measurable space $(\Omega, \mathcal{A})$ to the Borel line $(R, \mathcal{S})$ where R is the real line and $\mathcal{S}$ is the σ-field of its Borel sets. X induces the σ-field $\mathcal{B}(X) = X^{-1}(\mathcal{S})$ of events defined on it. Its (*probability*) *law* or *distribution* $\mathcal{L}(X)$ or F_X is defined by $F_X(S) = PX^{-1}(S)$, $S \in \mathcal{S}$, and $(R, \mathcal{S}, F_X)$ is its

state or *sample probability space.* The expectation EX of X is the integral $\int_\Omega X dP$—when this integral exists. Then its *conditional expectation* $E^{\mathfrak{B}}X$, or $E(X|\mathfrak{B})$, given a sub σ-field $\mathfrak{B}$ of events, exists and is a $\mathfrak{B}$-measurable function defined up to a P-equivalence by $\int_B E^{\mathfrak{B}}X\, dP = \int_B X\, dP$ for every $B \in \mathfrak{B}$. When $\mathfrak{B}$ is the smallest sub σ-field of events for which a function Y (numerical or not) on Ω is measurable, it is also denoted by $E(X|Y)$. When $X = I_A$, it is the *conditional probability* $P^{\mathfrak{B}}A$, or $P(A|\mathfrak{B})$ (or $P(A|Y)$) of the event A given $\mathfrak{B}$ (given Y).

A *random function* or (*stochastic*) *process* $X_T = (X_t, t \in T)$ is a family of random variables indexed by $t \in T$; however, later we shall use the term "process" in a more general sense and, meanwhile, we shall speak of random functions. The *sample space* of X_T is $(R^T, \mathcal{S}^T)$. Its *law* or *distribution* $\mathcal{L}(X_T)$ or F_{X_T} is defined by $F_{X_T} = PX_T^{-1}(S)$, $S \in \mathcal{S}^T$; according to the (Daniell-Kolmogorov) consistency theorem it is determined by the laws $\mathcal{L}(X_{t_1}, \ldots, X_{t_n})$ of all its finite sections. Its *sample probability space* is $(R^T, \mathcal{S}^T, F_{X_T})$. Its *expectation* EX_T is the family $(EX_t, t \in T)$—when all EX_t exist.

3. *Stochastic Properties.* *Stochastic* or *random analysis* is concerned with those properties of random functions X_T which can be expressed, directly or indirectly, in terms of their laws. Its main concern is with the behavior of *sample functions* $X_T(\omega) = (X_t(\omega), t \in T)$, $\omega \in \Omega$. In general, T is a *time set*—a subset of the real line with the usual topological, measurability, and algebraic structures. To simplify, we let T be $\{1, 2, \ldots\}$ or $[0, \infty)$, unless otherwise stated. In sample analysis there appear "sampling times" of various properties of sample functions. They are times $\tau(\omega)$ at which the sample functions $X_T(\omega)$ acquire these properties, for example, the first discontinuity time, the last exit time from a set S, and so on. If there is no such $\tau(\omega) \in T$, we set $\tau(\omega) = \infty$. When $\tau = (\tau(\omega), \omega \in \Omega)$ is a measurable function on Ω to $T \cup \{\infty\}$ it is a *random time.* Then it is an X_T-*time* if the event $[\tau \leqq t]$ is defined on $(X_s, s \leqq t)$ and $X_\tau = (X_{\tau(\omega)}(\omega), \omega \in \Omega)$ is a random variable; it is sometimes convenient to introduce an ideal state ∂ and set $X_{\tau(\omega)}(\omega) = \partial$ whenever $\tau(\omega) = \infty$.

4. *Independence and Dependence.* Probability theory comes into its own when random functions are given a "stochastic structure," that is, when their laws are given certain properties.

At the beginning was independence: random variables X_t, $t \in T$, are *independent* if $\mathfrak{L}(X_{t_1}, \ldots, X_{t_n}) = \Pi_{k=1}^{n} \mathfrak{L}(X_{t_k})$ for every finite subset $(t_1, \ldots, t_n)$ of T. The basic problem, and, until recently, the only one, was concerned with the behavior of sums of independent random variables. The laws of their sums are then *convolutions* or *compositions* of the laws of the summands. Equivalently, their *characteristic functions*, that is, the Fourier-Stieltjes transforms $\hat{F}$ of their laws F (with same affixes if any) are products of characteristic functions of the summands. The main lines of investigation are still the two following ones:

(a) *Convergence of laws* $F_n \to F$ in the sense of $\int_R g\, dF_n \to \int_R g\, dF$ for all $g \in C$, where C is the space of bounded continuous functions g on R. The main tool is the P. Lévy continuity theorem: F_n converges to some F if and only if $\lim \hat{F}_n$ exists and is continuous at the origin, and then $\lim \hat{F}_n = \hat{F}$.

(b) *Convergence to a random variable* of series and of arithmetic means of random variables in the sense of almost sure convergence ($\overset{\text{a.s.}}{\to}$), convergence in probability ($\overset{P}{\to}$), or convergence in the rth mean ($\overset{r}{\to}$).

The leit-motif of independence runs throughout the theory. Other stochastic structures are born from, and/or their investigation is guided by, the results obtained in the independence case. So far only a few sufficiently general and relatively simple structures are isolated.

Markov dependence is born from an attempt by Markov [57] to preserve the law of large numbers and the normal convergence, obtained in the Bernoulli case (coin tossing): $(X_t, t \in T)$ is *Markovian* if $P(X_{t_{n+1}} \in S \mid X_{t_1}, \ldots, X_{t_n}) = P(X_{t_{n+1}} \in S \mid X_{t_n})$ a.s. for every $S \in \mathcal{S}$ and every finite subset of points $t_1 < \cdots < t_n$ of T.

Martingales are born from an attempt by P. Lévy [47] to preserve properties of series, obtained in the case of independent summands: $(X_t, t \in T)$ when all EX_t exist, is called a (*sub*) *martingale* if

$$(X_{t_n} \leqq) X_{t_n} = E(X_{t_{n+1}} \mid X_{t_1}, \ldots, X_{t_n})$$

a.s. for every finite subset of points $t_1 < \cdots < t_n$ of T; sums $X_n = Y_1 + \cdots + Y_n$ of integrable random variables Y_k with $E(Y_{k+1} \mid Y_1, \ldots, Y_k) = 0$ a.s. (martingale differences) represent integrable martingales in the P. Lévy form.

Stationarity and second-order stationarity were transposed from the ergodic theory by Khintchine [37]: the random function $(X_t, t \in T)$ is *stationary* if $\mathcal{L}(X_{t_1}, \ldots, X_{t_n}) = \mathcal{L}(X_{t_1+h}, \ldots, X_{t_n+h})$ (for all finite subsets $(t_1, \ldots, t_n)$ and $(t_1 + h, \ldots, t_n + h)$ of T. The random function $(X_t, t \in T)$ with all $E|X_t|^2$ finite is *second-order r.f.* and it is *second-order stationary* if the covariance function $\Gamma(s, t) = EX(s)\bar{X}(t)$, $s, t \in T$, is stationary, that is, a function of $s - t$ only.

5. *Three Main Results.* Three propositions play a fundamental role in probability theory and in its applications, simultaneously as powerful tools and as sources for problems and directions of research.

The first proposition is the far-reaching extension by Doob [14, 15] of P. Lévy's zero-one law:

Martingales Convergence Theorem. Let $\{X_n\}$ be a (sub) martingale. If $\sup EX_n^+ < \infty$ then $X_n \xrightarrow{a.s.} X$. If the $|X_n|^r$ are uniformly integrable for some $r \geqq 1$ then, moreover, $X_n \xrightarrow{r} X$.

The second is essentially Birkhoff's [43] and von Neumann's [85] ergodic theorems, or the strong laws of large numbers, or

Stationarity Theorem. Let $\{X_n\}$ be stationary. If EX_1 exists then $n^{-1}(X_1 + \cdots + X_n) \xrightarrow{a.s.} E^{\mathcal{I}}X_1$, where $\mathcal{I}$ is the invariant σ-field of events defined on $\{X_n\}$ and invariant under translations of this sequence into $\{X_{n+m}\}$, $m = 1, 2, \ldots$. If $|X_1|^r$ is integrable for some $r \geqq 1$ then, moreover, $n^{-1}(X_1 + \cdots + X_n) \xrightarrow{r} E^{\mathcal{I}}X_1$.

If $\mathcal{I}$ *degenerates*, that is, consists only of events or probability 0 or 1, $\{X_n\}$ is said to be *ergodic* and then $E^{\mathcal{I}}X_1$ degenerates into EX_1.

The third proposition is relative to the "central limit problem"; that is, the problem of convergence of laws of sums $\Sigma_k X_{nk}$, $k = 1, \ldots, k_n \to \infty$, of independent summands X_{nk} (with laws F_{nk}) which are "uniformly asymptotically negligible (u.a.n.)," that is, with $\sup_k P[|X_{nk}| > \epsilon] \to 0$ for every $\epsilon > 0$. The solution of the problem yields the class of *infinitely decomposable* (*i.d.*) laws $\mathcal{L}(X)$—such that for every n there exist n independent and identically distributed random variables $Y_{n1}, \ldots, Y_{nn}$ with $\mathcal{L}(\Sigma_{k=1}^n Y_{nk}) = \mathcal{L}(X)$.

Infinite Decomposability Theorem:

KHINTCHINE [38]. *The class of limit laws of sums $\Sigma_k X_{nk}$ of u.a.n. independent summands coincides with the class of i.d. laws.*

P. LÉVY [46]. *I.d. laws F are characterized by a triple (α, β^2, L) such that*

$$\hat{F}(u) = \exp\left\{i\alpha u - \beta^2 \frac{u^2}{2} + \int_{-\infty}^{+\infty} \left(e^{iux} - 1 - \frac{iux}{1+x^2}\right) dL(x)\right\}$$

where α and β are real numbers, and where the P. Lévy function L is defined and nondecreasing on $(-\infty, 0)$ and on $(0, +\infty)$ with $L(+\infty) = 0$ and $\int_{-1}^{+1} x^2\, dL(x) < \infty$. (The symbol $\int_a^b$ stands for integration on $(a, 0) \cup (0, b)$.)

GNEDENKO [27]. $\mathfrak{L}(\Sigma_k X_{nk}) \to \mathfrak{L}(X)$ *necessarily i.d. (α, β^2, L), if and only if,*

(i) $\Sigma_k F_{nk}(I) \to L(I)$ *for every interval I in $(-\infty, 0)$ and in $(0, +\infty)$ whose endpoints are continuity points of L.*

(ii) $\lim_{\epsilon \to 0} \overline{\lim_n} \Sigma_k \sigma^2 X_{nk}{}^{\epsilon} = \beta^2$ *where $X_{nk}{}^{\epsilon} = |X_{nk}|$ or 0 according as $|X_{nk}| < \epsilon$ or $|X_{nk}| \geqq \epsilon$ and $\sigma^2 X_{nk}{}^{\epsilon} = E(X_{nk}{}^{\epsilon} - EX_{nk}{}^{\epsilon})^2$.*

6. *First Queries.* Research, within the traditional setup, is active in the convergence problem of (sub) martingales and stationary sequences, and it is extremely active in and about the central limit problem. Related queries will be found in Part II.

QUERIES 1. There are better sufficient conditions than

$$\sup EX_n{}^+ < \infty$$

for convergence of (sub) martingales $\{X_n\}$. Using their definition, but with conditioning by increasing σ-fields $\mathfrak{A}_n \supset \mathfrak{B}(X_1, \ldots, X_n)$, Snell's result [81] is that $\lim X_n(\omega)$ finite exists for almost all $\omega \in [\inf_k \sup_n E(X_n{}^+ | \mathfrak{A}_k) < \infty]$ (the fact that $\lim X_n(\omega) > -\infty$ was observed by Luther). Chow's result [8] is that $\lim X_n$ exists a.s. when $\int_{[\tau<\infty]} X_\tau{}^+ < \infty$ for every (X_n)-time τ. An immediate question arises: find a Snell form of Chow's result. (Since this talk was given this query was answered by Chazan.)

Another type of result which is better expressed in terms of martingale differences $Y_n = X_{n+1} - X_n$ is that $\lim X_n(\omega)$ exists and is finite for almost all $\omega \in [E(\sup Y_{n+1} | \mathfrak{A}_n) < \infty]$. It and others are particular cases of another of Chow's conditions but whose formulation is involved. The two types of results overlap but do

not contain one another. Question: Is there a sufficiently simple condition containing both types? The weaker the condition the more hope to answer a basic question: Find necessary and sufficient conditions for a.s. convergence of (sub) martingales; in particular, for a.s. convergence to a finite limit which would yield the three series criterion for series of independent summands.

In the case of series of independent summands, convergence in law implies convergence in probability which implies a.s. convergence—to a finite limit (the reverse implications are always true). Question: Under what conditions—to be automatically fulfilled in the above case—do these implications hold for (sub) martingales?

QUERIES 2. Riesz's method of proof for the stationarity theorem extends to nonstationary sequences and some results were obtained [54] without assuming a specific type of dependence. Question: Apply this method, or a related one, but to specific nonstationary types.

In the stationarity theorem, if the X_n are independent (and identically distributed) their sequence is ergodic and $E^{\mathcal{E}}X_1$ degenerates into EX_1. In fact, a converse result then exists: if $n^{-1}(X_1 + \cdots + X_n)$ converges a.s.—necessarily to a constant c, then c finite implies that EX_1 exists and equals c (Kolmogorov) while $c = +\infty\,(-\infty)$ implies that $EX_1^{+}(EX_1^{-}) = +\infty$ [55]. Question: Find necessary and sufficient conditions for a.s. convergence—necessarily to $+\infty$ or $-\infty$, when EX_1 does not exist. A partial result is known [12].

Little is known about the "speed" of convergence and especially of a.s. convergence, in the (sub) martingale and stationary cases. The problem is important and ought to be systematically investigated. A guideline may be provided by whatever is known about it in the independence case. Perhaps also the methods in [55, pp. 514–519] may help.

QUERIES 3. The existing proofs of Gnedenko's theorem are cumbersome. At least for didactic purposes, a less computational and technical proof would be of value. Perhaps it may help to note that condition (i) is equivalent to $\mathcal{L}(\min_k X_{nk}) \to \mathcal{L}(Y)$ and $\mathcal{L}(\max_k X_{nk}) \to \mathcal{L}(Z)$, with $F_Y(x) = 1 - e^{-L(x)}$ or 1 and $F_Z(x) = 0$ or $e^{L(x)}$, according as $x < 0$ or $x > 0$. Question: For what "intuitive" reasons do the two extrema of the summands determine the P. Lévy function L?

The central limit problem extends at once to a "central asymp-

totic problem." It is concerned with asymptotic equivalence of distribution F_n of sums $\Sigma_k X_{nk}$ and of i.d. distributions G_n, in the sense of

$$\int_R g\, dF_n - \int_R g\, dG_n \to 0 \qquad \text{for every } g \in C.$$

Sometimes this extension is unavoidable. For example, Doblin [13] showed that even in the case of independent and identically distributed random variables there may be no norming constants a_n and b_n such that the laws of normed sums

$$\frac{X_1 + \cdots + X_n}{b_n} - a_n$$

converge, except to a degenerate law (this is always possible). Yet the problem of asymptotic behavior remains. The idea of comparing laws of sums with laws of the desired limit type goes back to Lindeberg's method for normal convergence (extended in [51] to dependent summands and without finite moments). In fact, a scrutiny of the methods used in the central limit problem shows that the comparison method is already there. It has been isolated and, in part, systematically investigated [53]. Question: Transpose whatever is known about the central limit problem under various restrictions (normed sums, normed sums of identically distributed summands) to the asymptotic one.

The asymptotic problem comes in naturally when the summands are dependent, and then "weighted" comparison laws G_n appear. A conditional distribution F^c of a certain type, say normal or i.d., yields a weighted normal or i.d. distribution EF^c. This problem was investigated [55] without imposing a priori a specific type of dependence. Question: Investigate the central asymptotic problem for various specific types of dependence. Some partial results are known for the martingale case. For example, if the summands X_{nk} are martingale differences, then, under Liapounov's condition $\Sigma_k E|X_{nk}|^{2+\delta} \to 0 (\delta > 0)$, the laws $\mathfrak{L}(\Sigma_k X_{nk})$ are asymptotically equivalent to weighted normal ones [55]; similarly under Lindeberg's condition. Question: Investigate the asymptotic behavior of laws of martingale sequences $\{X_n\}$ and, in particular, of stationary ones. A partial result is known [3, 34] in the stationary ergodic case for square integrable martingales: $\mathfrak{L}(X_n/\sqrt{n})$ converges to a normal law.

QUERIES 4. The central limit problem can be immediately reformulated for random numbers k_n of summands. Besides a few generalities [53] and a result of Anscombe [1], the investigation is so far concerned only with the case of limit laws of normed sums of independent and identically distributed summands. The first results were obtained by Robbins [76] followed by Tate [83] in the case of random numbers k_n independent of the summands Y_k. Renyi [74] showed that, without this assumption, normal convergence of $\mathfrak{L}(\Sigma_{k=1}^{n} Y_k/\sqrt{n})$ implies that of $\mathfrak{L}(\Sigma_{k=1}^{n} Y_k/\sqrt{k_n})$ provided $k_n/n \xrightarrow{P} \xi$—a discrete random variable. Then the "discrete" restriction was eliminated [5]. All these results are but special cases of much more general ones of Wittenberg [87] who uses the comparison method by the (Kolmogorov-Smirnov) vertical distance $\rho(F_n, G_n) = \sup_x |F_n(x) - G_n(x)|$. It appears that this problem has now a satisfactory answer, but only in the case of laws of independent and identically distributed summands. It seems that the ground is prepared for an attack on the limit laws problem for normed sums, then for sequences of sums $\Sigma_k X_{nk}$. Practically nothing is known about other types of convergence. In fact, all the preceding queries in this section can be reformulated in terms of random numbers of summands. There appears then a number of problems and not always very difficult ones.

QUERIES 5. Convergence (and comparison) of laws in terms of the vertical distance is but one of possible types. A family of types may be defined as follows. If B is a subspace of bounded measurable functions, then F_n and G_n are *asymptotically B-equivalent* $(F_n - G_n \xrightarrow{B} 0)$ if $\int_R g dF_n - \int_R g dG_n \to 0$ for every $g \in B$. The usual convergence of laws obtains for $B = C$. Yet in Markov processes appear naturally various subspaces of C such as the subspaces of uniformly continuous functions, of those which vanish at infinity, and of those which vanish off compacts [55, pp. 622–630]. It seems potentially interesting to examine the whole central limit problem and its variants in terms of these types of convergence.

In fact, various types of convergence of laws and, especially, the vertical distance type are of value for the "central approximation problem": Find in terms of n how "close" is $\mathfrak{L}(\Sigma_k X_{nk})$ to its limit law or to a class of i.d. laws. This is an important problem, for limit theorems are not quite satisfactory when the "speed" of convergence is unknown. There is a very large number of partial

results, mostly in the case of independent and identically distributed summands. One direction of research began with Liapounov [49], Berry [2], Esseen [20], and others, and continues with Cramér [11], Petrov [68–69], and many others. There, the apparatus of characteristic functions is constantly used. Another direction, born recently, begins and continues with Prokhorov [70–71], Kolmogorov [41], Meshalkin [59], Rogozin, and Le Cam [45]. It is concerned primarily with vertical and P. Lévy distances of distributions F_n of sums of independent random variables Y_n, with common distribution F, from i.d. distributions. It uses direct probabilistic methods and exploits P. Lévy's concentration function. Some beautiful results are already known:

(1) If $s_n(G) = \sup_F \rho(F_n, G)$, there exist constants c, c' and i.d. distributions G_n, G_n' such that $s_n(G_n) \leqq cn^{-\frac{1}{2}}$ (Kolmogorov) and $s_n(G_n') \geqq c'n^{-\frac{1}{2}} (\log n)^{-4}$ (Meshalkin).
(2) If $P[X_n \neq 0] = p_n$ and $p = \sup p_n$, then $\sup_n \rho(F_n, G_n) \leqq 25p^{\frac{1}{3}}$, for i.d. distributions G_n given by $\hat{G}_n = \exp\{\sum_{k=1}^{n} (F_k - 1)\}$ (Le Cam). There is an overlap between Kolmogorov's and Le Cam's results but the connection is not clear and ought to be investigated. Also, it may be useful to introduce approximations by weighted i.d. laws.

QUERIES 6. A number of unsolved problems about characteristic functions arises from probabilistic considerations. For example, it would be useful to have a representation of the class of characteristic functions with a given modulus. One may begin with specific moduli. Note that if the X_n are independent random variables, then the laws of sums of random variables $\pm X_n$ have the same modulus. Another useful result would be an answer to the problem of representation of classes of characteristic functions with the same values in some neighborhood of the origin. In particular, when is the class a singleton? This problem is related to that of decomposition of laws and to the continuity theorem. A partial answer to the particular problem is known [42]. Important recent advances in the decomposability problem, due to Linnik, raise still other questions. Since all these problems, once stated, are essentially analytic, I shall be content here with referring to Linnik's book [50], to P. Levy [48], and to Lukacs [56].

Characteristic functions are, up to constant factors, those continuous functions of positive-definite type $\Gamma(s, t)$ which depend only

upon the difference $s - t$ of the arguments (Bochner). In turn, the class of all functions of positive-definite type coincides with that of covariance functions. It would be of value to second-order analysis to transpose known properties and unsolved problems for characteristic functions to covariance functions. For example, find conditions under which a continuous covariance function $\Gamma(s, t)$ is "harmonizable," that is, of the form

$$\Gamma(s, t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{i(ux-vy)}\, dd'\gamma(x, y),$$

where γ is a covariance. These conditions ought to be automatically satisfied when $\Gamma(s, t)$ is a function of $s - t$ only, so as to have an extension of the important Bochner theorem.

Functions of positive-definite type appear in various branches of mathematics (integral equations, reproducing kernels and functions of several complex variables, and so on). Their interpretation as covariances may yield new results or simplify existing proofs. For example, Schur's theorem on preservation of the positive-definite type under multiplications is immediate when independent random functions are used. Also there are mutual interactions between second-order random functions and reproducing kernels theories [55, p. 490] which ought to be investigated systematically. Only some very partial results are known [66].

II. DISCUSSION

The traditional setup is fast breaking down. The foundations remain but various traditional restrictions are being gradually removed. New mathematical problems and phenomena are being discovered. In fact, some applications of probability theory to mathematics and to physics are shaking its very foundations.

1. *Stochastic Structures.* Probability theory is concerned with stochastic structures. Yet, to quote Doob [16], "progress has certainly been slow in the development of new types." Our first task is to analyze the known types and the accretion of substructures.

Let $(\Omega, \mathfrak{A}, P)$ be our probability space and let the σ-fields under consideration be contained in $\mathfrak{A}$. To avoid technicalities, assume that every sub σ-field, say, $\mathfrak{B}$ contains the σ-ideal $\mathfrak{N}$ of P-null events. This can always be achieved; it suffices to replace $\mathfrak{B}$ by the σ-field generated by $\mathfrak{B}$ and $\mathfrak{N}$.

Let $(\mathcal{B}_t, t \in T)$ be a family of σ-fields. In phenomenological terms, $\mathcal{B}_t$ is the class of "observable outcomes" at time t. To every $t \in T$, we associate the *past* σ-field $\mathcal{A}_t = \bigvee_{r \leq t} \mathcal{B}_r$, the *future* σ-field $\mathcal{A}^t = \bigvee_{s \geq t} \mathcal{B}_s$ and the *tail* σ-*field* of the future $\mathcal{C} = \bigcap_{t \in T} \mathcal{A}^t$ (also the tail σ-field $\bigcap_{t \in T} \mathcal{A}_t$ of the past when T has no first element). If the $\mathcal{B}_t$ are nondecreasing in t so that $\mathcal{A}_t = \mathcal{B}_t$, we may and shall then write $\mathcal{A}_t$ in lieu of $\mathcal{B}_t$. We associate random times τ, say, to the σ-fields of the past: τ is an $(\mathcal{A}_t)$-*time* if $[\tau \leqq t] \in \mathcal{A}_t$, $t \in T$; all $t \in T$ are trivially such times.

A stochastic process $(X_t, \mathcal{B}_t, t \in T)$ is a family of σ-fields $\mathcal{B}_t$ with associated $\mathcal{B}_t$-measurable functions X_t. When $\mathcal{B}_t \equiv \mathcal{B}(X_t)$, that is, the X_t induce the $\mathcal{B}_t$, we need not and do not mention these σ-fields and thus we recover the traditional processes $(X_t, t \in T)$.

Total structures: stationarity. Families of σ-fields may be defined in terms of one σ-field $\mathcal{B}$ and an additive group or semigroup $(\theta_t, t \in T)$ of *translations*, that is, mappings of events into events which preserve complementations and countable unions hence intersections. Thus to $\mathcal{B}$ are associated its translates—the σ-fields $\mathcal{B}_t = \theta_t\mathcal{B}$; we may and do restrict $\mathcal{A}$ to $\bigvee_{t \in T} \theta_t\mathcal{B}$. The events which remain invariant under translations form the (θ_t)-*invariant* σ-*field* $\mathcal{I}$; note that it is contained in the tail σ-field $\mathcal{C}$. When $\mathcal{I}$ degenerates, the $\theta_t\mathcal{B}$ form an *ergodic* family.

Translates of sets yield translates of their indicators by $\theta_t I_B \equiv I_{\theta_t B}$ and, by extension, *translates* of measurable functions X into measurable functions θ_t by $[\theta_t X \in S] = \theta_t[X \in S]$, $s \in \mathcal{S}$. The family $(\theta_t\mathcal{B}, t \in T)$ is *stationary* if the probability P is (θ_t)-*invariant*, that is, $P\theta_t \equiv P$. Then the stochastic process $(\theta_t X, \theta_t\mathcal{B}, t \in T)$ is a *stationary* process.

Conditional independence. In the background of independence and of Markov dependence lies the concept of conditional independence of σ-fields (isolated in [55], because the random variables which figure in the traditional description of these types intervene only through their induced σ-fields.

Two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ of events are *conditionally independent* given a σ-field $\mathcal{C}$ if, $P^{\mathcal{C}}C_1C_2 = P^{\mathcal{C}}C_1 \cdot P^{\mathcal{C}}C_2$ a.s., $C_1 \in \mathcal{C}_1$, $C_2 \in \mathcal{C}_2$; if $\mathcal{C}_1$ and $\mathcal{C}_2$ are σ-fields then, equivalently, $P^{\mathcal{C}_1 \vee \mathcal{C}}C_2 = P^{\mathcal{C}}C_2$, $C_2 \in \mathcal{C}_2$ (or $P^{\mathcal{C}_2 \vee \mathcal{C}}C_1 = P^{\mathcal{C}}C_1$, $C_1 \in \mathcal{C}_1$), a.s. (since any pair of σ-fields is conditionally independent given $\mathcal{A}$, we may exclude this trivial case). In a σ-field $\mathcal{C}_1$ there is a largest such $\mathcal{C}$, namely $\mathcal{C}_1$ itself (whatever be $\mathcal{C}_2$). There is also a smallest one, namely the smallest

σ-field for which all $P^{\mathcal{C}_1}C_2$ are measurable—the *σ-field of conditional independence* of $\mathcal{C}_2$ upon $\mathcal{C}_1$; similarly for such σ-fields in a σ-field $\mathcal{C}_2$. Note that two families of σ-fields are conditionally independent given $\mathcal{C}$ if and only if it is so for the smallest σ-fields generated by each family.

This concept leads naturally to the following stochastic structure: the σ-fields $\mathcal{B}_t$, $t \in T$, are *($\mathcal{C}_t$)-independent* if $\mathcal{A}_t$ and $\mathcal{A}^t$ are conditionally independent given the σ-field $\mathcal{C}_t$ for every $t \in T$. If $\mathcal{C}_t \equiv \mathcal{B}_t$, we have *Markov dependent* σ-fields $\mathcal{B}_t$; note that for any family $(\mathcal{B}_t, t \in T)$ its σ-fields of conditional independence (of $\mathcal{A}^t$ upon $\mathcal{A}_t$) are Markov dependent. If $\mathcal{C}_t \equiv \mathcal{C}$, the $\mathcal{B}_t$ are *$\mathcal{C}$-independent.* In particular, if $\mathcal{C}$ degenerates, then the $\mathcal{B}_t$ are *independent.*

A stochastic process $(X_t, \mathcal{B}_t, t \in T)$ will be said to have any of the above structures if the $\mathcal{B}_t$ have it. Also, if the $\mathcal{B}_t$ are $\mathcal{C}$-independent and the X_t have same conditional law given $\mathcal{C}$, we may say that the process is *$\mathcal{C}$-exchangeable.* When $\mathcal{B}_t \equiv \mathcal{B}(X_t)$, Markov independence and exchangeability structures reduce to the traditional ones.

Among the stochastic structures based on conditional independence, $\mathcal{C}$-independence is, in many ways, the most direct extension of independence. Problems are immediate and not necessarily difficult, while the answers may be quite useful. For, transposition of results and problems in the independence case is obvious: apply the results to $\mathcal{C}$-conditioned distributions $F^{\mathcal{C}}$, then take expectations and weighted distributions $EF^{\mathcal{C}}$ appear. The first systematic work in this direction is that of Bühlmann [6]. He is concerned with exchangeable random variables X_n (centered at expectations) with finite variances. They are reduced to $\mathcal{C}$-independent and identically $\mathcal{C}$-conditionally distributed random variables, and weighted normal convergence follows. As a particular case, the following elegant necessary and sufficient condition for normal convergence of $(X_1 + \cdots + X_n)/\sqrt{n}$ is obtained: $n^{-1}(X_1 + \cdots + X_2) \xrightarrow{a.s.} 0$ and $n^{-1}(X_1^2 + \cdots + X_n^2) \xrightarrow{P} 1$. He also introduced and discussed random functions with "exchangeable increments." In a different direction, Freedman [22*a*] obtained important extensions of de Finetti theorem. Question: What about $\mathcal{C}$-exchangeable random variables without finite moments? Weighted stable laws will then appear. What about the whole central limit problem and its variants? What about other types of convergence?

Asymptotic structures: asymptotic independence. The main problem of random analysis is to describe the behavior of processes with an imposed structure as time t varies. It is frequently reduced to the problem of the asymptotic behavior of stochastic sequences (as $t \to \infty$), and then we shall assume, to simplify, that $T = \{1, 2, \ldots\}$ or $T = \{\ldots, -1, 0, 1, \ldots\}$, according to the context.

Asymptotic behavior ought to be determined by asymptotic stochastic properties. This leads to associating to stochastic structures their asymptotic types. The question is how to select those asymptotic properties of a structure so as to preserve the desired asymptotic results. The more of them we impose, the closer we may come to similar asymptotic behaviors. This field of research is at its beginning. It is mostly concerned with imposing upon a stationary process various versions of "asymptotic independence" so as to preserve normal convergence and sometimes a law of large numbers.

The weakest interpretation of asymptotic independence in the stationary case is ergodicity: the invariant σ-field degenerates. Then in the stationarity theorem the almost sure limit degenerates into EX_1, as in the case of independence. A slightly less weak interpretation is that the tail σ-field degenerates. This is equivalent to a "mixing" property (M_0): $\sup_{A \in \mathfrak{A}^t} |PAB - PAPB| \to 0$ for every $\mathfrak{B} \in \mathfrak{A}$. But to achieve the specific goals mentioned above, stronger mixing properties had to be introduced: There exists a function $c(t)$ decreasing to 0 from some t on, such that for every $A \in \mathfrak{A}_s$, $B \in \mathfrak{A}^{s+t}$

$$(M_1) \qquad |PAB - PAPB| \leqq c(t) \quad \text{or}$$

$$(M_2) \qquad |PAB - PAPB| \leqq c(t)PA \quad \text{or}$$

$$(M_3) \qquad |PAB - PAPB| \leqq c(t)PAPB.$$

The first was used by Rosenblatt [77] and the second by Rozanov [79] and Ibragimov [35], primarily to achieve normal convergence. The third one was used in [5] to obtain laws of large numbers. Some properties of the independent and identically distributed case were transposed. For example, Ibragimov found, under (M_1), necessary and sufficient conditions for normal convergence when the second moment is finite and showed that, under (M_2), the limit laws of normed sums are stable.

A different type of mixing was introduced by Renyi [73] and used to transpose limit laws theorems for normed sums of independent summands to various cases of dependence. A sequence $\{A_n\}$ of events is *mixing* if $PA_mA_n \to pPA_m$ $(0 < p < 1)$ for every m, equivalently, $PA_nA \to pPA$ for every $A \in \mathcal{A}$. It implies that $QA_n \to p$ for every P-continuous probability Q. Note that if there is a limit law for normed sums of independent summands X_n, then for every x the sequences of events $[X_n < x]$ have this mixing property. Sucheston [82] investigated such types of mixing and their direct consequences.

QUERY. Investigate systematically the influence of the various types of asymptotic independence for martingale and Markov structures, stationary then nonstationary. Consider the various types of convergence.

Asymptotic conditional independence. The importance of $\mathcal{C}$-independence becomes apparent when it is given an asymptotic form. The first systematic work in this direction is Cogburn's deep analysis of stationary sequences on $T = (\ldots, -1, 0, +1, \ldots)$ ([9] also [10]) containing as particular cases many results, new and known, under asymptotic independence. Among others, he uses the following asymptotic conditional independence given the invariant σ-field $\mathcal{I}$:

$$\left| \frac{1}{n} \sum_{k=0}^{n-1} P^{\mathcal{A}_0}\theta^k A - P^{\mathcal{I}}A \right| \leqq c(n)\downarrow 0$$

for all $A \in \mathcal{A}^0$. A number of results follows for the stationary Markov case, the central limit, and the ranking limit problems. Weighted i.d. laws become the standard ones, and degenerate into i.d. ones in the ergodic case. Question: Examine conditional forms for various mixing conditions mentioned above. What becomes of the known results under these conditional forms. Proceed then to the nonstationary case.

Asymptotic stationarity. From a phenomenological point of view, stationarity describes statistical equilibrium. But this is a "long run" property of phenomena—an asymptotic one. Thus, to provide more realistic models for phenomena, it is natural to consider "asymptotic stationarity." It may be given various forms, say, $P\theta_t \to \bar{P}$—where $\bar{P}$ is necessarily a (θ_t)-invariant probability, or $|P\theta_s - P\theta_t| \to 0$ as $s, t \to \infty$ under various restraints, say, $s = g(t)$

is a specific function of t and, in particular, $g(t)/t \to 0$ or $g(t)/t \to c > 0$ a constant to be specified or not. I know of no work in this direction. Question: Introduce and discuss forms of asymptotic stationarity according to phenomenological or mathematical considerations. What happens to stationarity results, especially to the fundamental stationarity theorem? Add ergodicity and various asymptotic and asymptotic conditional independence, conditions.

Martingales. The definition of (sub) martingales associated by Doob to σ-fields $\mathcal{A}_t = \vee_{s \leqq t} \mathcal{B}_s$ may be written $(X_s \leqq)\ X_s = E(X_t|\mathcal{A}_s)$ a.s. They are (sub) martingales in the traditional sense but possess more properties when $\mathcal{B}_t$ strictly contain $\mathcal{B}(X_t)$; when $\mathcal{B}_t = \mathcal{B}(X_t)$ so that $\mathcal{A}_t = \mathcal{B}(X_s, s \leqq t)$, then one recovers the traditional (sub) martingales $(X_t, t \in T)$. In its turn, this definition leads at once to an extension of the (sub) martingale structure: Let φ_t be consistent (nondecreasing) signed measures on nondecreasing $\mathcal{A}_t$, that is, $\varphi_s = \varphi_t(\varphi_s \leqq \varphi_t)$ for $s < t$. The family $(\varphi_t, \mathcal{A}_t, t \in T)$ may be called a martingale (submartingale). It reduces to the preceding type if and only if the φ_t are σ-finite and $P_{\mathcal{A}_t}$-continuous (with Radon-Nikodym derivatives X_t). However, in this new form, a simple property leads to new proofs and results: $\varphi_\infty = \lim_{t \to \infty} \varphi_t$ exists and is finitely additive on the field $\cup_{t \in T} \mathcal{A}_t$. To simplify, let $T = (1, 2, \ldots)$. (For nets T important results were obtained by Krickeberg, for example, [43] and by Chow [7].)

The consistency theorem becomes a martingales theorem on a specific probability space: For a martingale $(P_n, \mathcal{A}_n)$, where the P_n are probabilities on the σ-fields of cylinders in R^∞ with Borel bases in $R_1 \times \cdots \times R_n$, $P_\infty = \lim P_n$ is σ-additive on $\cup \mathcal{A}_n$ and hence determines a probability on $\vee \mathcal{A}_n$. This leads at once to the investigation of even traditional (sub) martingales $(X_n, n = 1, 2, \ldots)$ but on their sample spaces $(R^\infty, \mathcal{S}^\infty)$. For example, it provides another proof of a.s. convergence of integrable (sub) martingales with, moreover, $\varphi = \lim \varphi_n$ bounded and σ-additive if (and obviously only if) $\sup E|X_n| < \infty$ ([54], also [63] and [55], pp. 94, 402–407). In fact, in connection with this approach, there appear asymptotic behaviors of (to simplify) finite signed measures φ_n which yield a.s. convergence with no other structure imposed. For example, if there exists such a measure φ on $\vee \mathcal{A}_n$ with $|\varphi_n(A) - \varphi(A)| \leqq c(n)PA$ where $c(n) \to 0$, for every $A \in \mathcal{A}_n$ $(A \in \mathcal{A}^n)$, then their Radon-Nikodym derivatives (of the continuous parts) X_n converge a.s. to the Radon-Nikodym derivative of the restric-

tions to $\vee \mathfrak{A}_n$ ($\cap \mathfrak{A}^n$). This yields the a.s. convergence of (sub) martingales. Question: Would the same approach or a related one yield a.s. convergence for the stationarity theorem?

We return to (sub) martingales $(\varphi_n, \mathfrak{A}_n, n = 1, 2, \ldots)$. The question of a.s. convergence arises when $\varphi = \lim \varphi_n$ is only finitely additive. A satisfactory answer is known when φ is bounded [7]: $X_n \overset{\text{a.s.}}{\to} X_\infty$—the Radon-Nikodym derivative of the largest σ-additive part of φ as determined by the Hewitt-Yosida theorem [30]. Question: What happens to the Hewitt-Yosida theorem and to this proposition when φ is σ-finite, then when the φ_n are σ-finite? Question: What about a further modification with finitely additive φ_n and corresponding Radon-Nikodym derivatives à la Bochner-Phillips?

($\mathfrak{B}_t$)-classes of processes. A cycle of problems arises when, instead of one process, one considers classes of processes with random variables $X_t, Y_t, \ldots$ associated to a same family $(\mathfrak{B}_t, t \in T)$ of σ-fields or *($\mathfrak{B}_t$)-classes.*

QUERY. Investigate and find operations which preserve a given structure of a ($\mathfrak{B}_t$)-class. For example, the ($\mathfrak{B}_t$)-class of (sub) martingales is closed under formation of linear combinations $aX_t + bY_t$ ($a \geqq 0$, $b \geqq 0$ for submartingales), suprema, transformation into $g(X_t)$, $h(X_t)$ by continuous convex functions (and nondecreasing for submartingales), and so on. Are there other such operations? What about various limit operations? The question admits variants: One may have to change the probability measure to preserve the structure under an operation. For example, let $X_n \geqq 0$ with $EX_n \equiv 1$ form a martingale and the $Y_n \leqq 0$ form a submartingale, both associated to σ-fields $\mathfrak{A}_n$. Then $P_\infty A = \lim \int_A X_n$, $A \in \cup \mathfrak{A}_n$, exists but is not necessarily σ-additive, while its restrictions P_n to $\mathfrak{A}_n$ are probabilities. Yet [17] $(Y_k/X_k, k \leqq n)$ is a submartingale (martingale when the Y_n form a martingale and the $X_n > 0$) associated to the $\mathfrak{A}_k$ but with respect to P_n and not P. If P_∞ is σ-additive then $(Y_n/X_n, n = 1, 2, \ldots)$ is a submartingale (martingale when the $X_n > 0$) with respect to P_∞.

QUERY. Find subclasses of given ($\mathfrak{B}_t$)-classes which "generate" the whole class. The term "generate" may be given various meanings (if any), say, by means of certain operations. For example, in a different direction, under relatively weak conditions [60], submartingales are "generated" by martingales in the sense that they

can be decomposed into martingales and nondecreasing processes (which are always submartingales) associated to the same σ-fields.

QUERY. Find within a given $(\mathfrak{B}_t)$-class with a specified structure, the subclass which has also another type of structure. Some results in this direction are known. Let me refer to [70] where the problem of characterization of stationary processes $(\theta_n X, n = 0, \pm 1, \pm 2, \ldots)$ which can be obtained from independent and identically distributed random variables ξ_n by means of a Borel function g and the translations θ_n: $X_n = g(\theta_n\{\ldots, \xi_{-1}, \xi_0, \xi_1, \ldots\})$, is discussed. The query may be reversed: examine those structures which may appear under various specified operations in $(\mathfrak{B}_t)$-classes.

2. *Index Spaces.* The traditional index space T is a "time" set—a subspace of the real line with the usual structures inherited from it. At first glance, it permits us to describe the evolution of random phenomena. Yet, even from this phenomenological point of view, it may not be satisfactory. A space-time evolution may be required (quantum field theory). However, the most important fact is that the measurements are never instantaneous. They do not yield values at a fixed moment of time but a sort of average of the values in some neighborhood of this moment, say, $X(\varphi)(\omega) = \int X(\omega, t)\varphi(t)\,dt$ where the functions φ characterize the measuring apparatus.

Phenomenological as well as mathematical considerations lead to the following types of nontraditional index spaces T.

(1) T is a linear and, in general, locally convex topological space. Most frequently, T is a subspace of a multidimensional Euclidean space R^N (or of a Minkowski space). The corresponding "random fields X_T" theory is at its beginning. T may be a Hilbert space and, more generally, a Banach space; there is a fast-growing literature on this direction, pioneered by Mourier [62], Fortet [22], and others.

(2) T is a class of sets, especially, T is the Borel field on R or R^N. The corresponding "random measures" X_T appear already in second order harmonic analysis.

(3) T is a functional space such as spaces of continuous functions on subspaces of R or R^N. The most important ones are the spaces introduced by L. Schwartz and, in particular, the space $\mathcal{S}$ of infinitely differentiable functions φ with $|t|^n\varphi^{(k)}(t) \to 0$ as $|t| \to \infty$ for every integer n and every kth derivative $\varphi^{(k)}$, or the subspace $\mathcal{D}$ of such

functions vanishing off compacts. The main operations in $\mathfrak{D}$ which are transposable to random distributions are those which commute with shifts θ_h—defined by $\theta_h\varphi(t) = \varphi(t + h)$. Such are linear combinations of shifts and their limits, say, "differentiation" defined by $\varphi'(t) = \lim_{h\to 0} (\varphi(\theta_h t) - \varphi(t))/h$, "filtering" by K defined by $\int k(t - s)\varphi(s)\, ds$, and so on.

The random functions X_T are selected so as to reflect in some sense the structures of T. For linear topological index spaces they are linear and continuous. A *linear* random function is simply a linear mapping of a linear space T to the linear space $\mathfrak{R}$ of random variables on some probability space or to the space of their equivalence classes. When T is provided with a topology compatible with its linear structure, then various types of *continuous* linear X_T appear corresponding to the usual types of convergence in $\mathfrak{R}$. If T consists of differentiable functions φ, it is "natural" to define the derivative of $X(\varphi)$ by $X'(\varphi) = -X(\varphi')$ (this is suggested by formal differentiation of $\int X(t)\varphi(t)\, dt$ and integration by parts). The most important of such continuous linear random functions are random distributions.

A *random distribution* X is a linear mapping on $\mathfrak{D}$ to $\mathfrak{R}$ (or $\mathfrak{R}/\mathfrak{N}$) continuous, that is, $X(\varphi_n) \to X(\varphi)$ in pr or a.s. or in the rth mean in the sense of convergence in $\mathfrak{D}$: $\varphi_n \to \varphi$ if all the φ_n vanish off some compact and $\varphi_n^{(k)} \to \varphi^{(k)}$ uniformly for every $k = 0, 1, \ldots$. Similarly for "tempered random distributions"—with $\mathfrak{D}$ replaced by $\mathcal{S}$ with its type of convergence.

The mathematical importance of index spaces of distributions for probability theory is due to the fact that in the traditional setup sample functions are, in general, extremely unwieldly from the point of view of classical analysis: they are not differentiable, their Fourier transforms do not exist, and so on. Yet, once random functions are transformed into random distributions, the whole distribution theory becomes available and sample analysis becomes possible.

The corresponding "random distributions" theory was pioneered by Gelfand [26] and by Ito [32], and continued by Urbanik [84], Dudley [19], and others.

Random distributions theory is only at its beginning. So far, only independence and stationarity structures were transposed to random distributions. $X(\varphi_1)$ and $X(\varphi_2)$ are *independent* whenever

$\varphi_1\varphi_2 \equiv 0$. $X_{\mathfrak{D}}$ is *stationary* when $\mathfrak{L}(X(\varphi_1), \ldots, X(\varphi_n)) = \mathfrak{L}(X(\theta_h\varphi_1), \ldots, X(\theta_h\varphi_n))$ for every shift θ_h and every finite subset $(\varphi_1, \ldots, \varphi_n)$ of $\mathfrak{D}$, and $X_{\mathfrak{D}}$ is *second-order stationary* if $EX(\varphi) = EX(\theta_h\varphi)$, $EX(\varphi_1)X(\varphi_2) = EX(\theta_h\varphi_1)X(\theta_h\varphi_2)$ for every φ, every pair φ_1, φ_2, and every θ_h. There is need for a large scale systematic investigation: Transpose to random distributions, and investigate further, with all the analytical tools thus become available, the foregoing total and asymptotic structures, the traditional results and the foregoing queries.

There is another modification of index spaces which already proved its worth in Markov processes but is not yet used elsewhere. *The index spaces $T(\omega)$ vary with $\omega \in \Omega$.* In Markov theory they are time intervals with random right-end points. In Martin boundary theory, Hunt [31] introduced random left- and right-end points: $T(\omega) = [\alpha(\omega), \beta(\omega)]$ with α and β not necessarily finite. From a phenomenological point of view, such index spaces are natural for construction of models: a random phenomenon may be expected to have a random birth α and a random death β. There is need for a systematic investigation of the whole traditional setup, its results and problems, but with random index spaces.

3. *Abstract Sample Spaces.* The definition of a random variable takes into account the usual topology of the real line indirectly—through the construction of its σ-field of Borel sets. Yet, all it requires is a measurable sample space $(\mathfrak{X}, \mathcal{S})$. This leads at once to "abstract" *random variables* X, that is, functions X on $(\Omega, \mathcal{A})$ to $(\mathfrak{X}, \mathcal{S})$ such that $X^{-1}(\mathcal{S}) \subset \mathcal{A}$. From this point of view, any random function $X_T = (X_t, t \in T)$ (whether the X_t are numerical or abstract) is itself a random variable with sample space $(\mathfrak{X}^T, \mathcal{S}^T)$. However, this space has already some structure: it is a space of functions on T and sample analysis has content provided there is meaning to properties of these functions, such as continuity, measurability, and so on.

Topological and measurability structures of this functional space are, in general, the corresponding product structures of those imposed on $\mathfrak{X}$. (Most frequently, measurability in $\mathfrak{X}$ is introduced through its topology: $\mathcal{S}$ is the Borel field—the one generated by the topology, or $\mathcal{S}$ is one of the Baire fields—the smallest one for which continuous functions belonging to a subfamily of them are measurable.) Also X_T may be of second order and considered as a curve in

the Hilbert space $L_2(\Omega, \mathfrak{A}, P)$ of equivalence classes of random variables, or the $X_T(\omega)$ may be signed measures, and so on.

Whichever the case, we are faced with random variables which take values in some structured space such as a linear topological space or a compact or locally compact topological group. Their appearance may be due to some more or less "soft" generalization: given a proposition in the traditional setup, one may seek to impose structures to upon $\mathfrak{X}$ under which the proposition remains valid, in part or in toto. Or they may be due to a more or less "hard" generalization: in a mathematical problem with a given $\mathfrak{X}$ there appear beings or methods which recall probabilistic ones, and this leads to an investigation of the similarities but within the setup of the problem.

In fact, the first abstract random variables appeared in concrete problems: as far back as 1928, F. Perrin [6, 7] studied rotations of a sphere immersed in a fluid whose particles perform a Brownian motion. After n units of time, its total rotation is $\rho_n\rho_{n-1} \cdots \rho_1$, where the ρ's are independent random rotations. This raises the problem of (multiplicative) laws of large numbers for random variables with values in an orthogonal group (noncommutative). Ordered linear lattice chains of particles subject to continguous random forces are governed by linear difference equations, say, $u_n = \xi_n u_{n-1} + \eta_n u_{n-2}$ with (ξ_n, η_n) independent and identically distributed random vectors. Then $\begin{pmatrix} u_{n+1} \\ u_n \end{pmatrix} = X_n X_{n-1} \cdots X_1 \begin{pmatrix} u_1 \\ u_0 \end{pmatrix}$ where the $X_n = \begin{pmatrix} \xi_{n+1} & \eta_{n+1} \\ 1 & 0 \end{pmatrix}$ are random matrices, and the asymptotic behavior of u_n is governed by the matrix product $X_n X_{n-1} \cdots X_1$. Random phenomena governed by differential equations lead to random variables in a Lie group, and so on.

Thus, "hard" generalizations become unavoidable. They have to proceed from given structured sample spaces, and the problems are mostly guided by or are similar to the traditional ones. The need for "abstract" expectations, that is, abstract integrals arises. Quite a collection of them and procedures for formation of needed ones are available (see [58]). Also there is need for "abstract" conditional expectations or Radon-Nikodym type of derivatives. Here are available operator treatments begun by Moy [64].

At present, the problems under investigation are centered about independent and identically distributed random variables in the

direction of convergence of laws and the strong law of large numbers. They require a high degree of mathematical sophistication. Yet the subject is at its very beginning, and even transposition of various stochastic structures and traditional results remains to be done. Since there is a recent "exposé d'ensemble" by Grenander [28] of known results with an exhaustive bibliography, I shall be content to refer to it, adding only a few more recent papers I think of importance, by Furstenberg [23] and Gangoli [24–25] (there are many others as well).

I would like to mention here an unpublished result of Strassen which shows how new phenomena appear in a most classical problem: Let (T, d) be a compact metric space and let $X_T^{(n)}$ be independent and identically distributed square integrable random functions with continuous sample functions, that is, abstract random variables with values in the space C of continuous functions on T. Center the $X_t^{(n)}$ at their expectations: $EX_t^{(n)} \equiv 0$. It may be expected that the usual normal convergence holds, that is, $\mathcal{L}(\sum_{k=1}^n X_T^{(k)})/\sqrt{n})$ converges to a normal law in C. Yet, let H_δ be the δ-entropy of T, that is, the $\log_2$ of the smallest cardinal of δ-dense subsets S of $T(d(t, S) < \delta$ for every $t \in T)$. To simplify, assume that T is "not too big" in the sense that $\delta^2 H_\delta$ remains bounded as $\delta \downarrow 0$ and that $\beta(\delta) = \operatorname{esssup}_\omega \sup_{d(s,t)\leq\delta} |X_s^{(n)} - X_t^{(n)}| < \infty$. Then normal convergence holds if there exist constants $c > 0$ and $\alpha > 1$ such that $\beta(\delta) \leqq c/(\log 1/\delta)^\alpha \sqrt{H_\delta}$. However, given any T as described above, there exist $X_T^{(n)}$ with $\beta(\delta) \leqq 1/(\log 1/\delta)\sqrt{H_\delta}$ for which normal convergence does not hold, in fact, there is no convergence of laws!

4. *Extensions.* Sample spaces, selected in advance, are frequently too poor for some problems and have to be extended accordingly. I shall be content with a few examples.

In the traditional setup, the definition of random variables as taking values in the real line R breaks down at almost every step. The topologies for the space $\mathfrak{R}$ of random variables on $(\Omega, \mathcal{A}, P)$, and for its subspaces of random variables X with $E|X|^r < \infty$, corresponding to various types of convergence based upon P, are not separated. To make the limits unique one has and constantly does identify equivalent measurable functions. This introduces random variables which can take infinite values but with zero proba-

bility. Furthermore, $\mathfrak{R}$ is not closed under a.s. passages to the limit and this introduces random variables which can take infinite values with positive probability. Also sequences $\{F_n\}$ of laws are not compact in the sense of their traditional type of convergence yet become compact in the sense of the weaker type of convergence: $\int g\,dF_n \to \int g\,dG$ for every $g \in C_0$—the space of continuous functions vanishing at infinity. While G is still a measure, $G(R)$ may be less than one and $1 - G(R)$ may be thought of as that part of probability masses which escapes to infinity (or ceases to be observable). This compactness is in constant use in the traditional central limit problem. Thus, even within the traditional setup there is constant need to consider "extended" random variables—with the extended real line and its Borel field. There are many partial results for such random variables. While it would be little more than an exercise, it may be useful to examine systematically the traditional problems and results, in particular the central limit problem, for extended random variables.

Another example of necessary extensions, indicative of technicalities involved in Markov theory, is the Martin completion, first constructed by Martin for the classical theory of potential then transposed by Doob [18] (also Watanabe [86]) to Markov processes theory. To simplify, let $\mathfrak{X} = (1, 2, \ldots)$, but assigned its discrete topology. Let X_T be a stationary Markov function with transition probability $P_t(x, S)$, hence corresponding to a Markov semigroup $(P_t, t \in T)$ defined by $(P_t f)(x) = \int P_t(x, dy) f(y)$ (it is a series with our $\mathfrak{X}$) on the space of nonnegative measurable functions f on $\mathfrak{X}$. A function f is (P_t)-*invariant* (or "regular" or "harmonic"—according to the authors) when $P_t f \equiv f$, equivalently, when $f(X_t)$ form a martingale. The problem is to find "representations" of such functions and of the asymptotic behavior of the sample functions of X_T. To simplify, assume that $T = (1, 2, \ldots)$ and that all states are transient, that is, the "potential" $U(x, S) = \Sigma_t P_t(x, S)$ is finite for all finite S, and let μ be a probability with $\mu S > 0$ for all nonempty S. Form $K(x, y) = U(x, \{y\})/\int \mu(dz) U(z, \{y\})$ and complete $\mathfrak{X}$ with respect to the uniformity induced upon taking for Cauchy sequences $\{y_n\}$ either those with almost all y_n equal or those with almost all $y_n \in S$ for any given finite S and such that the $K(x, y_n)$ form a Cauchy sequence for every x. This is the Martin completion and the set $\mathfrak{M}$ of new points is the *Martin boundary*. $\mathfrak{M}$ is this boundary compact metric space with the measure μ on its

Borel field $\mathfrak{B}$. It permits us to describe the asymptotic behavior of the sample functions and also carries the representation of harmonic functions f: There exists a one-to-one correspondence between such f and the family of finite measures on $\mathfrak{B}$. Let us mention a provocative interpretation of this representation by Blackwell (unpublished): He establishes the same correspondence but with measures μ on the invariant (under shifts of coordinates) σ-field of Borel sets in $\mathfrak{X}$. Also the need for the Martin boundary proper disappears. Why? The connection between the Martin boundary and another—the Feller boundary—(and of their applications), deserves to be investigated further following Feldman [21].

Let us take this opportunity to mention several "exposés d'ensemble" of Markov theory and of its subfields and applications by Dynkin [19a], Ito and McKean [33], and Meyer [61].

As a last example, in a different direction, let me mention Kendall's use [36] of Sz-Nagy extension of Hilbert spaces. He observes that there exists frequently an excessive measure μ on the Borel field $\mathfrak{S}$ of R, that is, $\mu S \geqq \int \mu(dx) P_t(x, S)$. The space $L_2(\mu)$ on $\mathfrak{X}$ is a Hilbert space with inner product $(f, g) = \int f(x) g(x) \mu(dx)$, and the operators P_t on H form a semigroup of contractions [65]. H is imbedded into a larger Hilbert space on which the P_t become unitary operators with well-known integral representation. It suffices then to project them on H.

5. *Probability Spaces.* A probability space $(\Omega, \mathfrak{A}, P)$ is but a frame of reference. From a phenomenological point of view it is a mathematical fiction. Ideally, what one observes are values of random variables, that is, points of a sample space. In practice, observations yield only the fact that the values belong to some sets—events occur. All one may hope to achieve is a reasonably close approximation of the probability law of some random function X_T. The "observable" probability space is its sample probability space. Then why the underlying probability space? The reason is primarily that of mathematical convenience. The probability space is ubiquitous and protean. It may be taken to be any observable one. It may be selected to be rich enough to carry any desired family of probability laws with any structure. But more is possible, as follows.

Probability algebras. If one does not believe in miracles, that is, occurrences of any specified event of probability zero (the statisti-

cians and the physicists do not), or if one desires to avoid the technicalities such events entail (the probabilists may), these events may be eliminated: Identify events whose symmetric difference is of probability zero. Then $d(A, B) = P(A \Delta B)$ is a distance and the probability space becomes a complete metric algebra of events with only the zero element being of probability zero. The points $\omega \in \Omega$ disappear but then they have little meaning, if any. In fact, Kolmogorov [40], Halmos [29], and others, sought to replace probability spaces by probability algebras. However, according to the Loomis-Sikorski theorem, every probability algebra can be represented by a probability space modulo sets of probability zero. Thus, it appears that to start with underlying probability spaces or probability algebras is a matter of taste. Furthermore, the fact that we may use the Stone space representation provides us with a powerful tool, as follows.

Stone space representation. Given a probability space $(\Omega, \mathfrak{A}, P)$, the ($\sigma$-)field of events modulo null events is isomorphic to the field $\mathfrak{F}$ of clopen sets (simultaneously open and closed) in a compact Hausdorff space Ω' where $\mathfrak{F}$ is a base for the topology. To each Baire set A'—element of the σ-field $\mathfrak{A}'$ generated by this field—there corresponds a unique clopen image of an event A (which differs from A' by at most a meager set). Then $P'A' = PA$ (and P-null sets are meager sets and conversely) defines a probability P' on $\mathfrak{A}'$. To equivalence classes of random variables X on Ω there correspond equivalence classes of random variables X' on Ω' preserving L_p-norms, linearity, ordering, a.s. sequential convergence, and so on. Thus, whenever convenient, we can select $(\Omega', \mathfrak{A}', P')$ as our probability space. Let me mention a few advantages.

From a phenomenological point of view, a maximal trial, that is, observation of occurrences or nonoccurrences of all possible events, ought to determine the "state of nature" ω. In mathematical terms it would be a maximal filter of events—and the above representation has the required property.

In a different direction, a finitely additive non-negative set function on the field of clopen sets is vacuously σ-additive and thus determines a measure on $\mathfrak{A}'$. Also, one may form the metric algebra $(\mathfrak{A}, d)$ even when P is finitely additive and, by completing with respect to the metric, obtain σ-additivity. From this point of view, the quarrel between the proponents of finitely additive probability with σ-additivity proponents is of no mathematical importance.

On the other hand, this approach may be of importance for (sub) martingales, say $(X_n \geqq 0, \mathfrak{A}_n)$. The limit $\varphi_\infty = \lim \varphi_n$ is finitely additive on $\cup \mathfrak{A}_n$. If it is not σ-additive, use the foregoing methods to extend it to a measure. Then investigate the asymptotic behavior of the sequence X_n; similarly for general (sub) martingales $(\varphi_n, \mathfrak{A}_n)$.

I know of only one specific probabilistic instance where the Stone space representation was exploited. Doob [18a] proved the Rota [78a] limit theorem on bistochastic operators using a correspondence with Markov endomorphisms: Let T be a linear positive operator on $L_1(\Omega, \mathfrak{A}, P)$ to itself, with $T1 = 1$. To T there corresponds T' of same type on $L'(\Omega', \mathfrak{A}', P')$ defined by $T'(f') = (Tf)'$—the image of Tf in the above representation. Then T' is a Markov endomorphism in the sense that there exists a transition probability $P'(x', A')$ with $(T'f')(x') = \int P'(x', dy')f'(y')$ for almost every x'. This representation may perhaps be exploited for ergodic theory.

Conditional probability spaces. Applications of probability theory to integral geometry and to the theory of numbers seem to point out that the fundamental concept of one underlying probability P is too restrictive. There appear nonfinite measures μ on σ-fields and the problems require at once the use of families of "conditional" probabilities P_C defined by $P_C A = \mu AC/\mu C$ for the sets C with $\mu C < \infty$. To take care of this need, Renyi [72] introduced conditional probability spaces $(\Omega, \mathfrak{A}, P_C, \mathfrak{C})$ with conditional probabilities P_C on classes $\mathfrak{C} \subset \mathfrak{A}$ as the primary datum. They are axiomatized by

(i) P_C are probabilities on $\mathfrak{A}$ with $P_C C = 1$.
(ii) If A and B with $P_C B > 0$ are events and $BC \in \mathfrak{C}$ then $P_{BC} A = P_C AB/P_C B$.

Note that for $\mathfrak{C} = \mathfrak{A}$, the P_C are traditional conditional probabilities given C derived from one probability $P = P_\Omega$.

So far, only the traditional normal convergence and approximation together with the law of iterated logarithm were used in the applications to number theory. I shall be content to refer to the book by Koubilious [44] and the exhaustive literature quoted therein. It would be of interest to transpose systematically the various stochastic structures and their properties to Renyi setup and its particular cases.

Expectation spaces. To conclude, let me indicate rapidly new fissures which become visible when probability concepts are used in quantum theory. There appear two families of (abstract) random variables—"observables" X and "observables" Y—which are self-adjoint operators on a Hilbert space of "state functions" $\psi(x, y, z)$ and which commute within each family only. Their observable "values" are their eigenvalues and their probability laws in a state ψ are given by the densities $\bar{\psi}X\psi$ and $\bar{\psi}Y\psi$. However, there are no probability laws for pairs (X, Y) of noncommuting operators and the tranditional interpretation of all X and Y as the space of random variables on a common probability space disappears. Various ways to detour this obstacle were given by von Neumann, Segal, Wiener, and others. But one may try to "absorb" this obstacle by enlarging accordingly the traditional setup of probability theory, as follows [52].

Since in quantum theory the concept of expectation with its usual properties remains valid, this leads to the introduction of expectation (in lieu of probability) as a primary datum. One may use, say, the following "expectation spaces" setup. On a measurable space $(\mathfrak{X}, \mathfrak{A})$ define "abstract" random variables X to a common space, with the structures required below, either directly or, proceeding as in the numerical case, introduce "abstract" indicators I_A with $I_A{}^2 = I_A$ then simple or elementary "abstract" random variables and then general random variables X. *Formally,* $X = \int x(u)I(du)$ where $x(u)$ is a real or complex-valued $\mathfrak{A}$-measurable function and $I(du)$ is a "decomposition of the identity" indicator on $\mathfrak{A}$. There may be several such measurable spaces $(\mathfrak{X}, \mathfrak{A})$, $(\mathfrak{Y}, \mathfrak{B})$, and so on with corresponding random variables $X, Y, \ldots$ to a common range space. Introduce the expectation E: a positive linear operator on "positive" random variables $X, Y, Z, \ldots$ to the (extended) real line or, more generally, to the (extended) complex plane with $EI = 1$ and $E(\lim Z_n) = \lim EZ_n$ for $Z_n \uparrow Z$, and extend by linearity; *formally* $EX = \int x(u)EI(du)$. For example, *formally* the traditional setup corresponds to one family of random variables $x = \int x(u)I(du)$ with finite measurable functions $x(u)$ and a common resolution of the identity indicator $I(du)$ on $(\mathfrak{X}, \mathfrak{A})$ with correspondence $P(du) = EI(du)$ between the expectation operator and a probability measure P on $\mathfrak{A}$. Or $I(du)$ may be a resolution of the identity operator on a Hilbert space. The expectation operator E and a self-adjoint positive-definite operator

whose trace is one—a "probability operator P" correspond by EX = trace PX; in fact, there is a probability measure on $\mathfrak{A}$ determined by trace $(PI(du))$. Or there may be two resolutions $I(du)$ and $J(dv)$ of the identity operator on a Hilbert space and corresponding families of random variables $X = \int x(u)I(du)$ and $Y = \int y(v)J(dv)$. In general, the self-adjoint operators X and Y do not commute and $Z = XY$ are not self-adjoint. However, $\Re(XY) = (XY + YX)/2$ and $(XY) = (XY - YX)/i$ are self-adjoint and hence correspond to resolutions of the identity. There still exists a "probability operator" P such that, say, EZ = trace $(P\Re(XY)) + i$ trace $(P(XY))$. The operators with a common resolution $H(dw)$ of the identity are, in fact, defined on a common probability space with probability measure given by $EH(dw)$.

The foregoing "expectation spaces" setups may be of phenomenological interest and contain the traditional and the quantum theory setups. Thus, it would be of interest to investigate them systematically and search for transpositions of the various stochastic structures and properties.

REFERENCES

1. F. J. Anscombe, Large sample theory of sequential estimation, *Proc. Camb. Phil. Soc.*, **48** (1952).
2. A. Berry, The accuracy of the Gaussian approximation to the sum of independent variates, *Trans. Amer. Math. Soc.*, **49** (1941).
3. R. Billingsley, The Lindeberg-Lévy theorem for martingales, *Proc. Amer. Math. Soc.*, **12** (1961).
4. B. D. Birkhoff, Proof of the ergodic theorem, *Proc. Nat. Acad. Sci. USA*, **17** (1931).
5. J. R. Blum, D. L. Hanson, and J. I. Rosenblatt, On the central limit theorem for the sum of a random number of independent random variables, Technical Report No. 9, University of New Mexico (1962).

5a. J. R. Blum, D. L. Hanson, and L. H. Koopmans, On the strong law of large numbers for a class of stochastic processes, *Z. Wahrscheinlich Keitstheorie*, **2** (1963).

6. H. Bühlmann, Austauschbare stochastische Variabeln und ihre Grenzwertsaetze, *Univ. California Publ. Statist.*, **3** (1960).
7. Y. S. Chow, Martingales in a σ-finite measure space indexed by direct sets, *Trans. Amer. Math. Soc.*, **97** (1960).
8. Y. S. Chow, Convergence theorems of martingales, *Z. Wahrscheinlichkeitstheorie*, **1** (1963).
9. R. Cogburn, Asymptotic properties of stationary sequences, *Univ. California Publ. Statist.*, **3** (1960).

10. R. Cogburn, Conditional probability operators, *Ann. Math. Statist.*, **33** (1962).
11. H. Cramér, Sur un nouveau théorème-limite de la théorie des probabilités, *Actualités Sci. Ind., Paris* (1938).
12. C. Derman and H. Robbins, The strong law of large numbers when the first moment does not exist, *Proc. Nat. Acad. Sci. USA*, **41** (1955).
13. W. Doblin, Sur l'ensemble de puissances d'une loi de probabilité, *Ann. Ec. Norm. Sup.* (1947).
14. J. L. Doob, Regularity properties of certain families of chance variables, *Trans. Amer. Math. Soc.*, **47** (1940).
15. J. L. Doob, *Stochastic processes*, New York (1953) Wiley.
16. J. L. Doob, Present state and future prospects of stochastic process theory, *Proc. Intern. Cong. of Math.*, **3** (1954).
17. J. L. Doob, A Markov chain theorem, *The Harald Cramér Volume* (1959).
18. J. L. Doob, Discrete potential theory and boundaries, *J. Math. Mech.*, **8** (1959).
18a. J. L. Doob, A ratio operator limit theorem, *Z. Wahrscheinlichkeitstheorie*, **1** (1963).
19. R. M. Dudley, Fourier analysis of substationary processes with a finite moment, *Trans. Amer. Math. Soc.* (to appear).
19a. E. B. Dynkin, *Markov processes* (to appear), Springer, translated from Russian (1963).
20. G. Esseen, Fourier analysis of distribution functions, A mathematical study of the Laplace-Gaussian law, *Acta Math.* (1944).
21. J. Feldman, Feller and Martin boundaries for countable sets, *Illinois J. Math.*, **6** (1962).
22. R. Fortet, Normalverteilte Zufallselemente in Banachschen Räumen, Anwendungen auf zufällige Funktionen, *Tagung Wahrscheinl. Math. Stat., Berlin* (1954).
22a. D. A. Freedman, Invariants under mixing which generalize de Finetti's theorem, *Ann. Math. Stat.*, **33** (1962); **34** (1963).
23. H. Furstenberg, Noncommuting random products, *Trans. Amer. Math. Soc.*, **108** (1963).
24. R. Gangoli, Brownian transition functions and random walks on certain homogeneous spaces (to appear).
25. R. Gangoli, Isotropic infinitely divisible measures on symmetric spaces, *Acta Math.* (to appear).
26. I. M. Gelfand, General random processes (Russian), *Dokl. Akad. Nauk SSSR*, **100** (1955).
27. B. Gnedenko, On convergence of distribution laws of sums of independent random variables, *Dokl. Akad. Nauk SSSR*, **18** (1938).
28. U. Grenander, *Probabilities on algebraic structures*, New York (1963) Wiley.
29. P. R. Halmos, The foundations of probability, *Amer. Math. Monthly*, **51** (1944).
30. E. Hewitt and K. Yosida, Finitely additive measures, *Trans. Amer. Math. Soc.*, **72** (1952).

31. G. A. Hunt, Markov chains and Martin boundaries, *Illinois J. Math.*, **4** (1960).
32. K. Ito, Stationary random distributions, *Mem. Coll. Sci., Univ. of Kyoto*, **28** (1954).
33. K. Ito and H. McKean, *Diffusion theory* (to be published).
34. I. A. Ibragimov, Some limit theorems for stationary processes, *Teor. Veroyatnost. i Primenen*, **7** (1962).
35. I. A. Ibragimov, A central limit theorem for a class of dependent random variables, *Teor. Veroyatnost. i Primenen*, **8** (1963).
36. D. G. Kendall, Integral representations for Markov transition probabilities, *Bull. Amer. Math. Soc.*, **64** (1958).
37. A. Khintchine, Korrelationstheorie der stationäre stochastischen Prozesse, *Math. Ann.*, **109** (1934).
38. A. Khintchine, Zur Theorie der unbeschrönkt teibaren Verteilungsgesetze, *Rec. Math.*, **44** (1937).
39. A. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung* (1933).
40. A. Kolmogorov, Algèbres de Boole complètes, *Ann. Soc. Pol. Math.*, **20** (1948).
41. A. Kolmogorov, On the approximation of distributions of sums of independent summands by infinitely divisible distributions, *Sankhyā, Series A.*, **25** (1963).
42. M. G. Krein, On the problem of continuation of positive hermitian continuous functions, *Dokl. Akad. Nauk SSSR*, **26** (1940).
43. K. Krickeberg and C. Pauc, Martingales et derivation, *Bull. Soc. Math., France*, **91** (1963).
44. I. P. Koubilious, *Probabilistic methods in number theory* (in Russian) (1962).
45. L. LeCam, A note on the distribution of sums of independent random variables, *Proc. Nat. Acad. Sci. USA*, **50** (1963).
46. P. Lévy, Sur les intégrales dont les éléments sont des variables aléatoires indépendantes, *Ann. d. R. Scuola Norm. d. Pisa*, II, **3** (1934).
47. P. Lévy, Propriétés asymptotiques des sommes de variables aléatoires indépendantes on enchainées, *Journal de Math.*, **14** (1935).
48. P. Lévy, Quelques problèmes non résolus de la théorie des fonctions caractéristiques, *Ann. Mat. Pura Appl.*, **53** (1961).
49. A. Liapounov, Nouvelle forme du theorème sur la limite de la probabilité, *Mem. Acad. St. Petersburg*, **11** (1900); **12** (1901).
50. Y. V. Linnik, *Décompositions des lois de probabilités* (translated from Russian) (1962).
51. Michel Loève, Etude asymptotique des variables aléatoires liées, *J. de Math. P. et Appl.*, **24** (1945).
52. Michel Loève, Indicateurs abstraits et champs stochastiques, *C. R. Acad. Sci. Paris*, **228** (1949).
53. Michel Loève, On sets of probability laws and their limit elements, *Univ. California Publ. Statist.*, **1** (1950).
54. Michel Loève, On almost sure convergence, *Proc. Second Berkeley Symp. on Math. Stat. and Prob.* (1951).
55. Michel Loève, *Probability theory* (1955, 1960, 1963).
56. E. Lukacs, *Characteristic functions* (1960).

57. A. A. Markov, Investigation of a remarkable case of dependent trials (in Russian), *Izv. Akad. Nauk SSSR. Ser. Mat.*, **1** (1907).
58. E. J. McShane, Integrals devised for special purposes, *Bull. Amer. Math. Soc.*, **69** (1963).
59. L. D. Meshalkin, On the approximation of distributions of sums by infinitely divisible laws, *Theory of Prob. and Its Appl.*, **6** (1961).
60. P. A. Meyer, A decomposition theorem for supermartingales, *Illinois J. Math.*, **6** (1962).
61. P. A. Meyer, *Probability and potential* (to be published).
62. E. Mourier, Eléments aléatoires dans un espace de Banach, *Ann. Inst. H. Poincaré*, **13** (1953).
63. Shu-Teh Chen Moy, Measure extensions and the martingale convergence theorem, *Proc. Amer. Math. Soc.*, **4** (1953).
64. Shu-Teh Chen Moy, Characterizations of conditional expectation as a transformation on function spaces, *Pacific J. Math.*, **4** (1954).
65. E. Nelson, The adjoint Markov process, *Duke Math. J.*, **25** (1958).
66. E. Parzen, Regression analysis of continuous parameter time series, *Proc. Fourth Berkeley Symp. on Math. Stat. and Prob.*, I (1961).
67. F. Perrin, Etude mathématique du mouvement Brownien de rotation, *Ann. Sci. École Norm. Sup.* (1928).
68. V. V. Petrov, Extension of the Cramér's limit theorem for the non-identically distributed random variables, *Vestnik Leningrad. Univ.* (1953).
69. V. V. Petrov, Asymptotic analysis of some limit theorems of the probability theory, *Vestnik Leningrad. Univ.* (1961).
70. Yu. V. Prohorov, On sums of identically distributed random variables, *Dokl. Akad. Nauk SSSR*, **105** (1954).
71. Yu. V. Prohorov, On the uniform limit theorems of A. N. Kolmogorov, *Theory of Prob. and Its Appl.*, **5** (1960).
72. A. Renyi, On a new axiomatic theory of probability, *Acta Math. Acad. Sci. Hungar.*, **6** (1955).
73. A. Renyi, On mixing sequences of sets, *Acta Math. Acad. Sci. Hungar.*, **9** (1958).
74. A. Renyi, On the central limit theorem for the sum of a random number of independent random variables, *Acta Math. Acad. Sci. Hungar.*, **11** (1960).
75. F. Riesz, Sur la théorie ergodique, *Comm. Math. Helv.*, **17** (1945); **19** (1947).
76. H. E. Robbins, The asymptotic distribution of the sum of a random number of random variables, *Bull. Amer. Math. Soc.*, **54** (1948).
77. M. Rosenblatt, A central limit theorem and a strong mixing condition, *Proc. Nat. Acad. Sci. USA*, **42** (1956).
78. M. Rosenblatt, Independence and dependence, *Proc. Fourth Berkeley Symp. on Math. Stat. and Prob.*, II (1961).
78a. G. C. Rota, An "Alternierende Verfahren" for general positive operators, *Bull. Amer. Math. Soc.*, **68** (1962).
79. Yu. A. Rozanov, An application of the central limit theorem, *Proc. Fourth Berkeley Symp. on Math. Stat. and Prob.*, II (1961).
80. Z. Sidák, Integral representations for transition probabilities of Markov chains with a general state space, *Chech. Math. J.*, **12** (1962).

81. J. L. Snell, Applications of martingale system theorems, *Trans. Amer. Math. Soc.*, **73** (1952).
82. L. Sucheston, A zero-one property of mixing sequences of events, *Bull. Amer. Math. Soc.*, **68** (1962).
83. R. F. Tate, Contributions to the theory of random numbers of random variables, unpublished thesis, University of California, Berkeley (1952).
84. K. Urbanik, Stochastic processes whose realizations are generalized functions, *Teor. Veroyatnost. i Primenen*, **1** (1956).
85. J. von Neumann, Proof of the quasi-ergodic hypothesis, *Proc. Nat. Acad. Sci. USA*, **18** (1932).
86. T. Watanabe, On the theory of Martin boundaries induced by countable Markov processes, *Mem. Coll. Sci. Univ. Kyoto*, **A33**: 1 (1960).
87. H. Wittenberg, Limiting distributions of random sums of independent random variables, *Z. Wahrscheinlichkeitstheorie* (to appear).

6

Random Integrals of Differential Equations

J. Kampé de Fériet

INTRODUCTION

At the beginning of our brief account of the present status of research on random integrals of differential equations, we would like to point out that the mathematical theory, which has developed in the last three or four decades, has its motivation in physics; this physical background is quite apparent, even in the most abstract parts of the field. The search for a theoretical model of the Brownian motion of a particle and the turbulent flow of a fluid has quite naturally led to the definition of random integrals of ordinary or partial differential equations respectively; in this last domain an extension of the statistical mechanics of holonomic systems to continuous media has given a definite orientation to the whole development. This is the only excuse that an applied mathematician (who has devoted a large part of his research to fluid dynamics) can use in discussing these mathematical problems, which are becoming more and more abstract every day.[1]

[1] As an illustration of this trend, compare Paul Lévy [G.1.3] and E. B. Dynkin [G.1.2] exposition of the theory of Markov processes.

In an attempt to make this survey as self-contained as possible and for the benefit of readers not familiar with probability theory, we have condensed in an Appendix the definitions and properties of random functions which are extensively used in the research we are reviewing.

The asterisk (*) after a word indicates that the word is defined in the Appendix.

PART A

1. *Statistical Mechanics of Holonomic Systems.* It is an interesting historical fact that random integrals of ordinary differential equations were used extensively twenty years before the theory of random functions had its first mathematical formulation.[2] J. W. Gibbs [A.1.1] considers mechanical systems Σ which are holonomic, that is, their *state* can be defined by a finite number of coordinates: the k *configuration parameters:*

$$q_1, q_2, \ldots, q_k$$

and the k *conjugate moments:*

$$p_1, p_2, \ldots, p_k.$$

Thus, a state of Σ is represented by a point ω in the *phase space*[3] Ω; the motion of Σ is defined by the $2k$ Hamilton-Jacobi differential equations:

$$(1.1) \qquad \frac{dq_j}{dt} = \frac{\partial H}{\partial p_j} \qquad \frac{dp_j}{dt} = -\frac{\partial H}{\partial qj} \qquad j = 1, 2, \ldots k$$

where H (the *Hamiltonian*) is the *total energy* E of Σ (kinetic + potential). As a rule one supposes that Σ is *conservative*, that is,

[2] The fundamental paper of J. W. Gibbs was published in 1902; and the first random function* was introduced in a rigorous mathematical setting by N. Wiener in 1923; moreover, for a system in statistical equilibrium, the integrals are stationary random functions*; nevertheless it was not until 1934 that a clear mathematical definition of a stationary random function was given by A. N. Khintchine.

[3] As a *set*, Ω can be considered a subset of R^{2k}, but the natural metric on Ω is *not Euclidean.* The configuration space $(q_1; \ldots; q_k)$ is a Riemannian variety whose metric is defined by the kinetic energy of Σ; the phase space Ω itself has to be deduced from the configuration space by a construction described in [G.3.1, pp. 34–37].

H is a function of the (q_j, p_j) but not of the time t. Under some smoothness conditions for H (which are satisfied for all the systems under consideration in rational mechanics), one can prove an existence, uniqueness, and continuity theorem for (1.1.): between the initial state ω of Σ at $t = 0$ and its state ω_t at any time t, $-\infty < t < +\infty$, there is a one-one bicontinuous correspondence

$$\omega \leftrightarrow \omega_t$$

which can be expressed in the form

$$\omega_t = T_t\omega. \tag{1.2}$$

The T_t define an Abelian group of homeomorphisms of Ω onto itself, the property

$$T_{t+s} = T_tT_s = T_sT_t$$

expressing the *Huyghens principle* for Σ.

For a given ω, the set of all $T_t\omega$, where t ranges over the reals, defines a *trajectory* (or *orbit*) $\Gamma(\omega)$; any *phase function* $F(\omega)$ constant along each trajectory is a first integral of (1.1).

Now the statistical mechanics of Σ is based on the assumption that the initial state ω is chosen at random in Ω according to a given probability law P

$$\text{Prob}\,[\omega \in A] = P(A) \tag{1.3}$$

(when A belong to a prescribed σ-algebra of subsets of Ω). Σ being conservative, one is especially interested in statistical equilibrium; thus one considers only probability measures P invariant under the group of homeomorphisms T_t

$$P(T_tA) = P(A). \tag{1.4}$$

Thus any phase function $F(\omega)$ (P measurable) is a *stationary random function,** and all the statistical properties of the motion are independent of the time t, that is, the *mean* or *expectation** of any phase function $F(\omega) \in L(\Omega, P)$, $E(F)$ is independent of t. In particular, the $2k$ integrals of the equations (1.1) $[q_j(T_t\omega), p_j(T_t\omega)]$ are stationary random functions. Thus *the statistical mechanics of Gibbs really dealt with random integrals of the Hamilton-Jacobi equations.*

This fact is obscured by the language used by Gibbs and by most present-day physicists. They consider a "very large" ensemble of

N identical copies $\Sigma_1, \ldots, \Sigma_N$ of Σ, moving *independently* of each other (there is absolutely no interaction between two systems Σ_j and Σ_k! No collisions as in the kinetic theory!), and suppose that the number N_A of systems whose states are represented by a point ω in A is given by

$$\frac{N_A}{N} = P(A). \tag{1.4}$$

In this context statistical equilibrium simply means that N_A is independent of the time t. The cloud of points $\omega_1, \ldots, \omega_N$ distributed in Ω according to (1.4) might seem more concrete to the physicist than the probability $P(A)$, but it must be pointed out that it is only an approximation of the strong law of large numbers:

$$\text{Prob}\left[\lim_{N\to+\infty} \frac{N_A}{N} = P(A)\right] = 1, \tag{1.5}$$

the location of the points $\omega_1, \ldots, \omega_N$ in Ω being assimilated to a sequence of N independent trials. The idea of an *ensemble* introduced by Gibbs (1902) found a rigorous justification only in 1930, when the proof by A. N. Kolmogoroff of the strong law gave a precise mathematical meaning to (1.4),

Moreover, among all the admissible probability laws P, Gibbs chose a particular one, the *canonical distribution*. P is absolutely continuous with respect to the Lebesgue measure (based on $d\omega = dq_1 \ldots dq_k dp_1 \ldots dp_k$, which is invariant under T_t by Liouville's theorem) and one has

$$dP = Ke^{-\alpha E}\, d\omega. \tag{1.6}$$

Here also the justification of the canonical distribution had to wait several decades (1950): for a given value of the mean total energy $\bar{E}$, it is (1.6) which maximizes the *information entropy* of Σ [A.1.2].

Nevertheless, the introduction, though in an implicit and even obscure way, of the random integrals $q_j(T_t\omega)$, $p_j(T_t\omega)$, was a clear indication of the role that random integrals of differential equations were to play in many physical problems.

Let us point out that the stationary random functions $F(T_t\omega)$ introduced in statistical mechanics define, in a sense, the most general type of stationary random functions. J. L. Doob [G.1.1,

p. 509] proved that any stationary random function, defined on a probability space $(\Omega, \mathcal{S}, P)$, can be put into the form $F(T_t\omega)$, where T_t is a group of suitable measure-preserving set transformations (operating on the sets of the σ-algebra $\mathcal{S}$).

2. *Ordinary Differential Equations.* In Gibbs statistical mechanics randomness is introduced only through the initial conditions, but many dynamical problems give the impression that some kind of forces are acting at random. Thus, in the Hamilton-Jacobi equations, the Hamiltonian H has to be taken as a random function of the (p_j, q_j). In many cases this point of view has inspired fruitful research (for an example, see N. M. Krylov and N. N. Bogolyubov [A.2.14]).

One is thus led to the general problem:

Given a q-dimensional random vector field $\mathbf{F}(t, \mathbf{x}, \omega)$ *find a q-dimensional random vector* $\mathbf{x}(t,\omega)$ *satisfying the q differential equations:*

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(t, \mathbf{x}, \omega) \tag{2.1}$$

and taking a given initial value $\mathbf{x}(0, \omega) = \mathbf{x}_0$ *or even a random initial value* $\mathbf{x}(0, \omega) = \mathbf{x}_0(\omega)$.

If $q = 3$ and if $\mathbf{F} = \mathbf{U}$ represents a random velocity field, this problem is the exact mathematical formulation of the diffusion problem in the turbulent flow of a fluid: the velocity $\mathbf{U}$ being a known random function of the time t and the position $\mathbf{x}$ (Eulerian description of a flow), find the random trajectories of the particles of the fluid (Lagrangian description). In spite of its great importance one is still very far from the solution of this problem for a large class of random vector fields $\mathbf{F}$. In fluid dynamics one has to be content with partial results. For instance, if the mean vector $E[\mathbf{F}(t, \mathbf{x}, \omega)]$ and the covariance tensor* $E[\mathbf{F}(t, \mathbf{x}, \omega)\mathbf{F}(s, \mathbf{y}, \omega)]$ are known, what can be said about the mean vector $E[\mathbf{x}(t, \omega)]$ and the covariance tensor $E[\mathbf{x}(t, \omega)\mathbf{x}(s, \omega)]$? Even in this limited field, very few results have been obtained so far (see, from the viewpoint of fluid dynamics, G. K. Batchelor [A.2.1, A.2.2]).

The only general results known are concerned with stationary random functions and differential equations with constant coefficients. K. Karhunen [A.2.13] has proved the following theorem.

If the equation $a_n\xi^n + \cdots + a_1\xi + a_0 = 0$ has no purely

imaginary root $\xi = i\kappa$, and if $y(t, \omega)$ is a normal stationary random function with an absolutely continuous spectrum

$$d\mathcal{E}_y(\kappa) = \mathcal{E}_y'(\kappa)\, d\kappa,$$

then the differential equation

$$a_n \frac{d^n x}{dt^n} + \cdots + a_1 \frac{dx}{dt} + a_0 x = y(t, \omega) \tag{2.2}$$

has a unique solution which is a normal stationary random function with the absolutely continuous spectrum*

$$d\mathcal{E}_x(\kappa) = \frac{\mathcal{E}_y'(\kappa)\, d\kappa}{|a_n(i\kappa)^n + \cdots + a_1(i\kappa) + a_0|^2}. \tag{2.3}$$

The normal stationary random function $y(t, \omega)$ being represented by a Wiener integral*

$$y(t, \omega) = \int_{-\infty}^{+\infty} e^{it\kappa} \sqrt{\mathcal{E}_y'(\kappa)}\, dW(\kappa, \omega), \tag{2.4}$$

then the random integral of (2.2) can be represented by

$$x(t, \omega) = \int_{-\infty}^{+\infty} \frac{e^{it\kappa} \sqrt{\mathcal{E}_y'(\kappa)}\, dW(\kappa, \omega)}{a_n(i\kappa)^n + \cdots + a_1(i\kappa) + a_0}. \tag{2.5}$$

Both integrals are interpreted by Karhunen as limits in the mean; but one can sharpen the result by using criteria for the almost sure existence and continuity of the derivatives up to the order n of a normal stationary random function. In fact, if $\mathcal{E}_y(\kappa)$ satisfies

$$\int_{-\infty}^{+\infty} [\log (1 + |\kappa|)]^{1+\epsilon}\, d\mathcal{E}_y(\kappa) < +\infty \qquad \epsilon > 0,$$

then (G. Hunt [A.2.12]) $y(t, \omega)$ is almost surely continuous.

But now (2.3) implies:

$$\int_{-\infty}^{+\infty} \kappa^{2n} [\log (1 + |\kappa|)]^{1+\epsilon}\, d\mathcal{E}_x(\kappa) < +\infty.$$

Thus, due to the criterion[4] of Yu. K. Belyaev [A.2.3] $x(t, \omega)$ is almost surely continuous and has almost surely continuous derivatives up

[4] In Belyaev's formula ϵ is omitted; this contradicts G. Hunt's criterion in the particular case $n = 0$. J. Delporte, by a completely different method, has established an equivalent criterion, but he finds the exponent $(1 + \epsilon)$ as was to be expected [A.2.11].

to the order n. Thus one can interpret (2.4) and (2.5) (see J. Kampé de Fériet [C.2, C.3]) as almost surely existing and (2.5) defines the random integral of (2.2) for almost all samples.

For nonlinear equations we know of few results. One interesting case was considered by R. H. Cameron [A.2.9, A.2.10]

$$\frac{dx}{dt} = F[t, x(t, \omega) + W(t, \omega)] \qquad x(0, \omega) = 0. \tag{2.6}$$

The author gives two sets of conditions on $F(t, u)$ in the strip $\{0 \leq t \leq 1, -\infty < u < +\infty\}$ sufficient for the almost sure existence of a random integral $x(t, \omega)$ in the interval $0 \leq t \leq 1$. Let us observe that if one puts

$$y(t, \omega) = x(t, \omega) + W(t, \omega)$$

then (2.6) is equivalent to the integral equation

$$y(t, \omega) - \int_0^t F[s, y(s, \omega)]\, ds = W(t, \omega).$$

It is interesting to note that for $F(t, u) = -u^2$, the conditions are not satisfied and it has been proved by D. A. Woodward [A.2.9, p. 840] that, in fact, for almost all ω $y(t, \omega)$ does not exist.

Of course the general equation (2.1) can also be written as a Volterra integral equation

$$\mathbf{x}(t, \omega) = \mathbf{x}_0(0, \omega) + \int_0^t F[s, \mathbf{x}(s, \omega), \omega]\, ds$$

which shows that the research of A. T. Bharucha-Reid [A.2.6, A.2.7, A.2.8] on random solutions of integral equations can be useful in this field.

3. *Stochastic Differential Equations.* In sharp contrast to the scanty material of the preceding section, when we come to the stochastic differential equations we find an exuberant wealth of material which will keep us busy for the next four sections.

The initial nucleus was a mathematical model of Brownian motion. In 1828, a botanist, Robert Brown, observed that very small solid particles suspended in a liquid perform completely erratic motions. From further observations it became increasingly clear that these irregular motions were an outward manifestation of the motion of the molecules of the liquid. The turning point came in 1905 when A. Einstein proved that the projection of the

particle, let us say on the Ox axis, did follow a normal law with mean $\bar{x} = 0$ and variance $\overline{x^2} = \sigma^2 t$. Seldom has a formula so simple had such far-reaching consequences and become so popular. In fact, it made possible the determination of the Avogadro Number (an accomplishment for which Jean Perrin was awarded the Nobel Prize in 1926). In 1916 M. Smoluchowski gave a confirmation of Einstein's ideas and improved on several points in his theory. Then, in 1923, N. Wiener [A.3.15] gave a rigorous definition[5] of a random function $W(t, \omega)$, which manifests the most typical features observed in the Brownian motion:

(a) The displacement Δx of the particle during the interval $[t, t + \tau]$ being produced (if τ is large enough) by a very large number of molecular impacts must, as a sum of a large number of independent random variables, follow the normal law (central limit theorem). This is true if $\Delta x = \sigma[W(t + \tau, \omega) - W(t, \omega)]$, even if τ is small (this last circumstance is possibly annoying for the physicist!).

(b) The displacement during a given time interval depends only on the molecular impulses acting during this interval. Thus the displacements $\Delta_1 x$ and $\Delta_2 x$ during two nonoverlapping time intervals $[t_1, t_1 + \tau_1]$, $[t_2, t_2 + \tau_2]$ must be independent. In fact, this property is satisfied by $\Delta x_1 = \sigma[W(t_1 + \tau_1, \omega) - W(t_1, \omega)]$ and $\Delta x_2 = \sigma[W(t_2 + \tau_2, \omega) - W(t_2, \omega)]$.

Thus the random function

$$x = \sigma W(t, \omega) \tag{3.1}$$

is a possible choice for theoretical model of the Brownian motion of a particle on the Ox axis; obviously, the Einstein formulas are verified by the displacement $W(t + \tau, \omega) - W(t, \omega)$ during any interval of time $[t, t + \tau]$.

But some properties of the random function W are rather puzzling for the physicist. First, $W(t, \omega)$ almost never has a derivative with respect to t. Thus the velocity of the particle is not defined. Second, $W(t, \omega)$ is almost never of bounded variation; the length of the path during any finite interval of time is infinite.

[5] In fact, this function was already considered in 1900 by L. Bachelier [A.3.1.] in quite a different context. He was looking for a mathematical model of the fluctuations of the stock exchange! Many of the characteristic features of $W(t, \omega)$ were discovered by L. Bachelier, but without rigorous proofs. The theory of probability was in its infancy at that time and one had to be satisfied with rather vague heuristic considerations.

In order to avoid these paradoxes, another approach has been suggested by Ornstein and Uhlenbeck [A.3.14] and later by S. Bernstein [A.3.3]. The position of the particle (starting from 0 at time 0) is now defined by

$$x(t, \omega) = \sigma \int_0^t e^{-\beta\xi} W(e^{2\beta\xi}, \omega)\, d\xi. \qquad \beta > 0. \tag{3.2}$$

The velocity $u(t, \omega)$ is well defined at any time t by

$$u(t, \omega) = \sigma e^{-\beta t} W(e^{2\beta t}, \omega). \tag{3.3}$$

This random function is normal, almost surely continuous in any finite interval, stationary with mean

$$E[u(t, \omega)] = 0 \tag{3.4}$$

and covariance

$$E[u(t, \omega)u(s, \omega)] = \sigma^2 e^{-\beta|t-s|} \tag{3.5}$$

and spectrum

$$d\mathcal{E}_u = \frac{\sigma^2}{\pi} \frac{2\beta\, d\kappa}{\kappa^2 + \beta^2}; \tag{3.6}$$

moreover, it is a *Markov process.**

In order to legitimize the model (3.2) for the Brownian motion, let us write the equation of motion of the particle on the Ox axis

$$\frac{du}{dt} = -\beta u + y(t, \omega). \tag{3.7}$$

The influence of the liquid is split into two parts: a systematic force, the friction $-\beta u$ (Stokes law); a fluctuating part, the random force produced by the molecular impacts. If we take

$$y(t, \omega \mid \tau) = \frac{\sigma}{\tau} [W(t + \tau, \omega) - W(t, \omega)],$$

then the force y is a normal stationary random function and (3.7) is a particular case of (2.2). Thus there is a unique normal stationary random integral $u(t, \omega)$; the spectrum of $y\ (t, \omega \mid \tau)$ being

$$d\mathcal{E}_y = \frac{2\sigma^2}{\pi} \frac{(1 - \cos \kappa\tau)}{\kappa^2\tau^2}\, d\kappa$$

the spectrum of $u\ (t, \omega)$ is given by

$$d\mathcal{E}_u = \frac{2\sigma^2}{\pi} \frac{(1 - \cos \kappa\tau)}{(\kappa^2 + \beta^2)\kappa^2\tau^2}\, d\kappa. \tag{3.8}$$

Now if $\tau \to 0$, then $d\mathcal{E}_y \to \frac{\sigma^2}{\pi} d\kappa$, which does not define a spectrum [this means that $y(t, \omega \mid \tau)$ does not tend to a limit; W has no derivative], but

$$d\mathcal{E}_u \to \frac{\sigma^2}{\pi} \frac{d\kappa}{\kappa^2 + \beta^2}$$

which is precisely the spectrum of the velocity in the Ornstein-Uhlenbeck process. This limiting property of equation (3.7) is usually expressed by the symbolic formula

$$du = -\beta u dt + \sigma dW \tag{3.9}$$

called Langevin's equation. One must carefully avoid dividing by dt because dW/dt has no meaning. Formula (3.9) is a symbolic substitute for the integral equation

$$u(t, \omega) - u(o, \omega) = -\beta \int_0^t u(s, \omega)\, ds + \sigma W(t, \omega) \tag{3.10}$$

One finds in J. L. Doob [A.3.4] a rigorous mathematical theory of Langevin's equation,[6] from this point of view.

The generalization of Langevin's equation introduced the following *standard type* of stochastic differential equation:

$$dx = m(t, x)\, dt + \sigma(t, x)\, dW. \tag{3.11}$$

The meaning of the function m and σ is given (J. L. Doob [G.1.1, p. 275]) by the limits of two conditional expectations:

$$\lim_{h \downarrow 0} E\left[\frac{x(t + h, \omega) - x(t, \omega)}{h} \,\middle|\, x(t, \omega) = \xi\right] = m(t, \xi)$$

$$\lim_{h \downarrow 0} E\left[\frac{[x(t + h, \omega) - x(t, \omega)]^2}{h} \,\middle|\, x(t, \omega) = \xi\right] = \sigma(t, \xi)^2.$$

The fact that, when $x(t, \omega)$ is known, the mean and variance of the

[6] S. Bernstein [A.3.3] defines Langevin's equation as the limit of a finite difference equation. M. Kac [A.3.8] gives a theory of Langevin's equation and of a more general equation (with external forces, such as gravity, acting on the particle) as the limit of a random walk.

increment $x(t + h, \omega) - x(t, \omega)$ are both of the order of h explains why P. Lévy [G.1.3] always writes the stochastic differential equation (3.11) in the following way:

$$dx = m(t, x)\, dt + \sigma(t, x) X \sqrt{dt},$$

X being, for each t, a normal random variable with mean 0 and variance 1. Under some regularity conditions for m and σ, the solution of (3.11) is a Markov process almost surely continuous. The effective construction of Markov processes from stochastic differential equations has been an outstanding contribution of Paul Lévy's to the theory of random functions.

Here let us make an amusing semantic point: the denominations "random functions" and "stochastic (or random) processes" are both used in the literature; the choice of one or the other by an author is motivated by his intuitive approach to the theory.

Let us consider a given set of functions $x(t, \omega)$ depending on a random parameter. If we pick a particular sample, at random, we have at our disposal a function which is already known in its whole interval of definition: "random function" clearly expresses this idea. If we start from a stochastic differential equation, we assume that we proceed to build the function step by step. When we know it up to a given time t, the equation defines the increment Δx for the next step; the word "process" expresses perfectly this knowledge progressing continuously from t to $t + \Delta t$ (see [G.1.3,] pp. 27 and 28]). This explains the reason why Paul Lévy has always remained faithful to "les processus stochastiques" and why we, having been introduced to the field through statistical mechanics (a set of given trajectories), are inclined towards "les fonctions aléatoires."[7]

The main result on the standard type of stochastic differential equations is due to K. Ito [A.3.7]. He makes the following hypotheses:

(H_1) $m(t, \xi)$ and $\sigma(t, \xi)$ are Baire functions in the strip

$$\{(t, \xi): 0 \leq t \leq 1, -\infty < \xi < +\infty\}.$$

[7] Let us point out that we are not always consistent; for instance, we say "Markov process," bowing to common usage. To be self-consistent in our terminology, we should call them "random functions of the Markov type" (as does, for instance, R. Fortet [A.4.2]); we hope that the reader will pardon this venial sin.

(H_2) In this strip

$$|m(t, \xi)| \leq K(1 + \xi^2)^{1/2}$$
$$0 \leq \sigma(t, \xi) \leq K(1 + \xi^2)^{1/2}$$
K constant.

(H_3) In the strip there hold the uniform Lipschitz conditions

$$|m(t, \xi_2) - m(t, \xi_1)| \leq K|\xi_2 - \xi_1|$$
$$|\sigma(t, \xi_2) - \sigma(t, \xi_1)| \leq K|\xi_2 - \xi_1|.$$

Then replacing (3.11) by the integral equation:

$$x(t, \omega) - x(0, \omega) = \int_0^t m[t, x(s, \omega)]\, ds + \int_0^t \sigma[t, x(s, \omega)]\, dW(s, \omega)$$

(the second integral is a new type of stochastic integral, whose meaning has been previously defined by K. Ito [A.3.5]), K. Ito gives an effective construction (by successive approximations) of the unique solution $x(t, \omega)$, which is an almost surely continuous Markov process if $t \in [0, 1]$. Moreover, for each $t \in [0, 1]$, $x(t, \omega) - x(0, \omega)$ is independent of all the increments $W(s + h, \omega) - W(s, \omega)$, if $t < s < s + h \leq 1$.

In a recent paper A. V. Sokorod [A.3.13] has split the existence and uniqueness theorem into two parts; for the existence of an almost surely continuous Markov process in [0, 1] the conditions

(H_1') $m(t, \xi)$ and $\sigma(t, \xi)$ are continuous in the strip
(H_2') as above
are sufficient.

4. *Transition Probability and Diffusion Equations.* In the preceding section the construction of the Markov process was based on the direct consideration of the stochastic differential equations, but there is another approach which has inspired a large amount of research. This other method defines $x(t, \omega)$ by the transition probability* $P(s, \xi, t, \eta)$. It is supposed that P satisfies a kind of Lindeberg condition

$$\int_{|\eta - \xi| > \epsilon} dP(s, \xi, t, \eta) = \text{Prob}\,[|x(t, \omega) - x(s, \omega)| > \epsilon | x(s, \omega) = \xi]$$
$$= 0(t - s) \qquad 0 \leq s < t$$

for every $\epsilon > 0$. The functions $m(t, \xi)$ and $\sigma(t, \xi)$ are now defined by

$$m(t, \xi) = \lim_{(t-s) \downarrow 0} \frac{1}{t - s} \int_{|\eta - \xi| > \epsilon} (\eta - \xi)\, d_\eta P(s, \xi, t, \eta)$$

$$\sigma(t, \xi)^2 = \lim_{(t-s) \downarrow 0} \frac{1}{t - s} \int_{|\eta - \xi| > \epsilon} (\eta - \xi)^2\, d_\eta P(s, \xi, t, \eta)$$

In 1931 A. N. Kolmogoroff [A.4.5] proved that under suitable regularity conditions, the transition probability satisfies the *backward diffusion equation* (with respect to the past conditions s, ξ)

$$\frac{\partial P}{\partial s} + m(s, \xi) \frac{\partial P}{\partial \xi} + \frac{1}{2} \sigma(s, \xi)^2 \frac{\partial^2 P}{\partial \xi^2} = 0; \tag{4.1}$$

and the *forward equation* (with respect to the present conditions t, η)

$$\frac{\partial p}{\partial t} + \frac{\partial}{\partial \eta} [m(t, \eta) p] - \frac{1}{2} \frac{\partial^2}{\partial \eta^2} [\sigma(t, \eta)^2 p] = 0, \tag{4.2}$$

where $p(s, \xi, x, t) = \dfrac{\partial}{\partial \eta} P(s, \xi, x, t)$.

The first is a parabolic equation with respect to the variables (s, ξ) for $s \leq t$; (t, η) enters only from the initial condition

$$\begin{aligned} P(t, \xi, t, \eta) &= 1 \qquad \xi < \eta \\ &= 0 \qquad \xi > \eta. \end{aligned} \tag{4.3}$$

The second is also a parabolic equation with respect to (t, η) for $t \geq s$; (s, ξ) enters only from a similar condition

$$\begin{aligned} P(s, \xi, s, \eta) &= 1 \qquad \eta > \xi \\ &= 0 \qquad \eta < \xi \end{aligned} \tag{4.4}$$

In particular cases the forward equation had been known a long time previously. L. Bachelier [A.3.1] used the heat equation ($m = 0, \sigma = 1$) extensively in connection with $W(t, \omega)$. A. Einstein, M. Smoluchowski, Fokker, M. Planck, and others, considered the case $m =$ constant, $\sigma =$ constant.

W. Feller [A.4.1] proved, under suitable regularity conditions for $m(t, \xi)$ and $\sigma(t, \xi)$, that $P(s, \xi, t, \eta)$ is uniquely determined by equations (4.1) and (4.2) and the initial conditions (4.3) and (4.4).

R. Fortet [A.4.2] has established the almost sure continuity of the Markov process defined by $P(s, \xi, t, \eta)$ when Feller's conditions are satisfied; more precisely he has shown that, almost surely

$$|x(t'', \omega) - x(t', \omega)| < 2C \left[|t'' - t'| \log \frac{1}{|t'' - t'|} \right]^{1/2}$$

for all $C > 1$, provided $|t'' - t'| < \delta(\omega)$, the random variable δ being independent of t' and t''.

R. Fuchs [A.4.3, pp. 192–193] refined the properties of the modulus of continuity of $x(t, \omega)$ and proved a proposition which is, in a sense, the converse of the preceding result, where starting from given $m(t, \xi)$, $\sigma(t, \xi)$, one concludes to the existence and continuity of $x(t, \omega)$. R. Fuchs proves that for every Markov process almost surely globally continuous (globally meaning a somewhat stronger kind of continuity, which he defines) the functions $m(t, \xi)$, $\sigma(t, \xi)$ exist for almost all t in $[0, 1]$.

5. *Semigroups and Infinitely Divisible Laws.* In the stochastic differential equation (3.11), the random impulse is represented by dW; in doing so we keep the spirit of Langevin's equation for Brownian motion. But it is easy to observe that, for the construction step by step of the Markov process $x(t, \omega)$, the most important feature of $W(t, \omega)$ is not the normal law but the independence of its increments. This points to a much more general type of stochastic differential equation,

$$dx = m(t, x)\, dt + \sigma(t, x)\, dV, \tag{5.1}$$

where the random impulse dV is now generated by the most general random function $V(t, \omega)$ with independent increments. $W(t, \omega)$ is simply a particular case of this class of random functions. Another well known but quite different example is the Poisson process (the samples are step-functions).

P. Lévy (see, for example, [G.1.3, pp. 147–203]) has proved that, after the subtraction of a suitable given function (not random), the function $V(t, \omega)$ can always be reduced to a random function with independent increments having no fixed singularities (centering of the process, according to modern nomenclature). For this last type the increment $V(t, \omega) - V(s, \omega)$ must follow an *infinitely divisible law*. The general structure of these laws was discovered by P. Lévy;

see [G.1.3, p. 161], where the general expression of the characteristic function of an infinitely divisible law is given explicitly.

For the separable version of $V(t, \omega)$ almost all samples are bounded on every finite interval, $a \leq t \leq b$ and their discontinuities are jumps.

The fundamental existence and uniqueness theorem has been proved by K. Ito [A.3.7] explicitly for the generalized standard type (5.1). Under the conditions (H_1), (H_2), and (H_3) the solution $x(t, \omega)$ is still a Markov process, but the samples are almost surely discontinuous.

As in the particular case of Section 4, the determination of the transition probability has been the subject of extensive research; W. Feller [A.4.1] has proved that the backward diffusion equation (4.1) becomes an integro-differential equation which reduces to a partial differential equation only in the continuous case.

The composition law of the transition probability has opened the door to the application of the theory of *semigroups*. Since the fundamental paper of K. Yosida (1949) [A.5.8] considerable work has been done and this has become one of the most fashionable spots in the field of Markov processes (W. Feller (1954–1955) [A.5.3, A.5.4], E. Hille and R. S. Phillips (1957) [A.5.6], A. Neveu (1958) [A.5.7], E. B. Dynkin (1959) [A.5.1, A.5.2] to mention only a few.)

To give one example, we refer the reader to pp. 648–660 of E. Hille and R. S. Phillips [A.5.6] where semigroup theory is applied to the particular case of Markov processes homogeneous in space and time, that is,

$$P(s, \xi, t, \eta) = P(t - s, \eta - \xi). \tag{5.2}$$

Through the Chapman-Kolmogorov* equation, written as a convolution,

$$P(t_1 + t_2, x) = P(t_1, x) * P(t_2, x),$$

the transition probability is interpreted as a semigroup of transformations acting in $L(-\infty, +\infty)$. The choice of this function space is motivated by the formula

$$P(t, \cdot)F(x) = \int_{-\infty}^{+\infty} F(x - \eta)\, d_\eta P(t, \eta)$$

giving the probability law of the process at the time t, when the initial probability is defined by $F(x)$.

Using the general methods of semigroup theory, the general expression of the characteristic function $\varphi(t, z)$ corresponding to $P(t, x)$ is found, giving a new proof of Paul Lévy's main result on indefinitely divisible laws:

$$(5.3)\quad \frac{1}{t}\log\varphi(t,z) = imz - \frac{1}{2}\sigma^2 z^2 + \int_{-\infty}^{+\infty}\left[e^{izx} - 1 - \frac{izx}{1+x^2}\right]\frac{1+x^2}{x^2}\,dG_0(x),$$

where m and σ are real constants ($\sigma > 0$) and G_0 is bounded, never decreasing and continuous at $x = 0$ (the case $G_0 = 0$ corresponds to the normal law). Moreover, for any function f such that f, f' are absolutely continuous and $f, f', f'' \in L(-\infty, +\infty)$ the *infinitesimal generator* A (in the strong topology) of the semigroup of transformations $P(t, \cdot)f$ has the following expression:

$$(5.4)\quad [Af](x) = -mf'(x) + \frac{\sigma^2}{2}f''(x) + \int_{-\infty}^{+\infty}\left[f(x-y) - f(x) - \frac{yf'(x)}{1+y^2}\right]\frac{1+y^2}{y^2}\,dG_0(x).$$

This gives a new interpretation of the forward diffusion equation (4.2) in the particular case $G_0 = 0$ and shows clearly that in the general case of (5.1) the forward diffusion equation has to be replaced by the integro-differential equation

$$\frac{\partial f}{\partial t} = [Af](x).$$

The great importance of the infinitesimal generator (strong or weak) is due to the fact that it completely determines the Markov process [A.5.1, p. 31].

Most of the results obtained in this way are of great importance for the theory of Markov processes, but are not within the scope of the present survey, having no direct bearing on the stochastic differential equations.

6. *Generalizations.* As a generalization of (5.1) one can consider the system of q stochastic differential equations

$$(6.1)\quad d\mathbf{x} = \mathbf{m}(t, \mathbf{x})\,dt + S(t, \mathbf{x})\,d\mathbf{V}$$

where $\mathbf{x}$ and $\mathbf{V}$ are q-dimensional random vectors

$$\mathbf{x} = [x_1(t, \omega), \ldots, x_q(t, \omega)]$$
$$\mathbf{V} = [V_1(t, \omega), \ldots, V_q(t, \omega)];$$

the q random functions $V_j(t, \omega)$ having independent increments; the vector $\mathbf{m}(t, \xi)$ and the matrix $S(t, \xi)$ being defined on the $(q + 1)$-dimensional set $0 \leq t \leq 1$, $\xi \in R^q$.

Both methods—direct construction of the q-dimensional Markov process $\mathbf{x}(t, \omega)$ and determination of the transition probabilities from the (generalized) Kolmogorov's equations—can be used without any essentially new difficulty (see [A.6.1]).

A very interesting approach from the physical viewpoint considers stochastic differential equations with vanishing diffusion, that is, the system

$$d\mathbf{x} = \mathbf{m}(t, \mathbf{x})\, dt + \epsilon S(t, \mathbf{x})\, dW$$

where the parameter ϵ is small. The method of Yu. N. Blagoveshchenskii [A.6.1] is based on an asymptotic expansion of the vector $\mathbf{x}(t, \omega)$ in a power series in ϵ; the first term being the solution of the system of q ordinary differential equations corresponding to $\epsilon = 0$. Curiously enough, the first approximation is Markovian, but not the kth for $k \geq 2$.

I. I. Gihman [A.6.3] has broadened the meaning of (6.1), dropping the hypothesis of independence for the increments $V_j(t, \omega) - V_j(s, \omega)$. He also supposes that the dV_j depend not only on the time t, but also on some space variable (eventually even infinite dimensional). This last hypothesis is motivated by the consideration of dynamical systems under the influence of random disturbances acting differently at different points of the phase space Ω. The integral of the system (6.1) is thus defined (in the spirit of S. Bernstein [A.3.3.]) as the limit (in the sense of convergence in mean square) of a sequence of functionals defined by recursive formulas.

A completely different type of generalization is obtained in the following way: In the preceding sections the Brownian motion function and, more generally, the random functions with independent increments took their values in R first and, now in this section, in R^q; but the Brownian motion has been extended successfully to non-Euclidean space (see P. Lévy [G.1.3, pp. 194–203]). In par-

ticular, the general structure of the functions with independent increments is completely determined for the circle (pp. 186–191) and for the sphere (pp. 202–203). Thus a new type of stochastic differential equations can be considered where the random function $x(t, \omega)$ now takes its value in a non-Euclidean space.

Along this line K. Yosida [A.6.4] has extended the theory to the forward and backward diffusion equations for the transition probabilities in a connected domain of an m-dimensional orientable Riemannian space.

PART B

7. *Statistical Theory of Turbulence.* It has been known, at least since the beginning of the nineteenth century, that the flow of water from a reservoir through a long circular pipe can have two completely different aspects; if the flow is slow enough, the trajectories are parallel straight lines and the velocity at a given point remains steady and varies smoothly from one point to another. If, however, the flow is fast enough, the trajectories become extremely complicated and the velocity displays variations in space and time which change unpredictably. Moreover, it has been found (H. Poiseuille, 1851) that in the first case, that of laminar flow, the experimental measurements are in very good agreement with the theoretical results deduced from the Navier-Stokes equations (1822–1846) for incompressible viscous fluids. On the contrary, in a turbulent flow the loss of pressure along the pipe can be 100 or even 1000 times greater than the value predicted from the equations. It was recognized as early as 1872 by J. Boussinesq that the chaotic nature of the flow was such that ignoring the fluctuations one must deal only with *averages.* This remark is the beginning of the statistical theory of turbulence. O. Reynolds (1895) went further; by averaging the Navier-Stokes equations, he gave the equations connecting the mean values of the components u_j of the velocity field and the mean values of the products $u_j u_k$ (Reynolds tensor). It must be borne in mind that for J. Boussinesq and even today for most physicists, the averages are always taken with respect to time t (at a fixed point $\mathbf{x}$) on a given realization of the flow (a sample of the ensemble of flows; see F. N. Frenkiel [B.7.4] for a careful comparison of the different types of averages commonly used). It was only much later, around 1930, that it was understood that, as in J. W.

Gibbs' statistical mechanics of holonomic systems, one must compute the averages over a very large ensemble of possible realizations of the turbulent flow. In the language of probability we say that the velocity field is a *random velocity field* $\mathbf{u}(\mathbf{x}, t, \omega)$, the phase space Ω and the probability measure P being suitably chosen.

At the beginning (for a general historical background see [B.7.1 or B.7.2]) the infant theory was confronted with many problems: jets, wakes, boundary layer, and so on. But, about 1930, a problem slowly emerged whose mathematical formulation did not seem to be hopeless; that is, the motion of a fluid when the walls are so far apart that their influence is negligible (mathematically the fluid fills the whole space). The Navier-Stokes equations being invariant for the translations $\mathbf{x} \to \mathbf{x} + \mathbf{h}$, it is clear that one can expect the most simple statistical results if one supposes that all *statistical averages are invariant under the translations* or, more precisely, that the velocity $\mathbf{u}(\mathbf{x}, t, \omega)$ is, at each time t, *a stationary random vector field* with respect to $\mathbf{x}$ (*homogeneous turbulence*).

One of the first approaches to the construction of such stationary random vector fields has been described by N. Wiener, under the picturesque name of "*homogeneous chaos*" [B.7.20].

The main idea consists in the representation of a random variable $F(\omega)$ belonging to $L^2(\Omega)$ as the limit in quadratic mean of a sum

$$\sum_1^\nu P_n(\omega)$$

where $P_n(\omega)$—called an homogeneous polynomial chaos—is itself a sum of products of Hermite polynomials

$$A_{n_1, \ldots, n_k} \prod_{n_1 + \cdots + n_k = n} H_{n_j}[W(I_j, \omega)]$$

W being the Wiener function and the intervals $I_1, \ldots, I_k$ being nonoverlapping. Wiener introduced his theory explicitly in view of statistical problems and with special reference to homogeneous turbulence. Nevertheless, it has never been exploited in full and only recently has it again attracted attention [B.8.10]. The main point of Wiener's chaos is that one does not have to make any a priori hypothesis about the probability law of $F(\omega)$. The coefficients $A_{n_1, \ldots, n_k}$ have to be computed in each particular problem and, when they are known, they determine the probability law.

Of course, only the first homogeneous polynomial chaos follows the normal law. The sum of the other polynomials ($n \geq 2$) measures the discrepancy of the actual law from the normal.

So far the theory has developed along another line, concentrating on the moments

$$E[u_{j_1}(\mathbf{x}_1, t, \omega) \ \ldots \ u_{j_n}(\mathbf{x}_n, t, \omega)] \tag{7.1}$$

which are the more easily obtained experimental data. Here, the *correlation tensor*

$$R_{j,k}(\mathbf{x} - \mathbf{y}) = E[u_j(\mathbf{x}, t, \omega)u_k(\mathbf{y}, t, \omega)] \tag{7.2}$$

and its Fourier-transform the *spectral tensor* $S_{j,k}(\kappa)$ play the fundamental role.[8] In [B.7.3] G. Birkhoff and J. Kampé de Fériet tried to put the theory in a rigorous measure-theoretical setting. We consider the linear topological[9] space Λ of all q-dimensional *vector* fields $\mathbf{u}(\mathbf{x})$ defined on R^p which are square integrable on any compact set D

$$E_D = \int_D |\mathbf{u}(\mathbf{x})|^2 \, dm(\mathbf{x}) < +\infty \qquad m = \text{Lebesgue measure on } R^p.$$

We call a measure on a linear topological space *regular* if it is the Lebesgue completion of a measure defined on the Borel sets. A regular measure P on Λ is admissible if the kinetic energy E_D has finite expectation for any compact D. We prove that any regular measure is strictly separable and metrically separable and establish the connection between admissible measures and random vector fields measurable with respect to $P \times m$. We introduce the spectral matrix measure $S_{j,k}(\kappa)$ and the correlation matrix $R_{j,k}(\mathbf{h})$ and prove that the correlation matrix of each admissible homogeneous vector field is associated with one and only one spectral matrix

[8] In the one-dimensional case the correlation and the energy spectrum were introduced by G. I. Taylor; generalized harmonic analysis [B.7.19] was explicitly suggested to N. Wiener by these ideas of G. I. Taylor. The definition of the correlation tensor is due to von Kármán and Howarth (1938). The spectral tensor was introduced in 1948 by G. K. Batchelor and J. Kampé de Fériet. Our paper [B.7.9] was an attempt to put the physical intuitions of Taylor in the mathematical frame of the newborn theory of random functions (1939).

[9] Λ is topologized by the condition

$$\mathbf{u}_n(\mathbf{x}) \to \mathbf{u}(\mathbf{x}) \qquad \text{means} \int_D |\mathbf{u}_n(\mathbf{x}) - \mathbf{u}(\mathbf{x})|^2 \, dm(\mathbf{x}) \to 0$$

for all compact D.

measure and conversely. We establish a uniqueness theorem proving that there is a one-one correspondence between the class of spectral matrix measures satisfying the symmetry condition $S_{j,k}(\kappa) = S_{j,k}(-\kappa)$ and the class of *normal* admissible random vector fields. Finally, we prove that any normal admissible stationary random vector field having reflection symmetry, whose spectrum is absolutely continuous, is *metrically transitive*, that is, *the statistical averages and the averages computed in the space R^p are equal with probability one.*

From this resumé, compressing into a few lines forty years of research, we conclude that the kinematics of homogeneous turbulence is now in good mathematical standing. However, this is only the beginning of a statistical theory of turbulence. According to the views of statistical mechanics, we know the phase space Ω, that is, the set of vector fields which can represent at a given time t, the state of the system (defined by the velocity field); now we have to determine the trajectory in Ω, starting from an arbitrary initial state.

Let us admit that any motion, even turbulent, of an incompressible viscous fluid must satisfy the Navier-Stokes equations (this is a very controversial hypothesis and, to our knowledge, no crucial argument pro or con has definitely closed the discussion). Then the answer is straightforward, that is, the evolution must be described by a random integral of the Navier-Stokes equations.

This sounds clear and simple, but unfortunately the theory of random integrals of Navier-Stokes equations still belongs to the realm of wishful thinking!

We know very little about the integrals of these nonlinear equations. Until recently, only elementary integrals (with special symmetry properties) were obtained. In 1933 the first existence and uniqueness theorems were proved by J. Leray. In the last ten years the field has become very active; and great progress has been made through the work of D. Graffi, E. Hopf, J. L. Lions, G. Prodi, I. A. Wolska-Bochanek, and O. A. Ladyshenskaya. Even if we are now able to escape the pitfall of branching or exploding integrals, nevertheless one fact remains; a set of integrals is never a linear space, which makes the discovery of significant topologies and measures in the phase space Ω very difficult.

As long as our statistical theory of turbulence is fundamentally based on Navier-Stokes equations, the success of our quest depends

on the progress of our ideas on the random integrals of these equations. The road leading to the construction of these integrals seeming barred, most research has been oriented toward the moments (7.1) and their evolution with time t.

One uses a method which basically goes back to O. Reynolds: Starting from the Navier-Stokes equations written for the velocity field $\mathbf{u}(\mathbf{x}, t, \omega)$, one takes their scalar product with the vector field $\mathbf{u}(\mathbf{y}, t, \omega)$, and averaging (with respect to ω), one obtains the values of the derivatives

$$\frac{\partial}{\partial t} R_{j,k}(\mathbf{x} - \mathbf{y})$$

(the so called Kármán's equations for the propagation of correlation). These time derivatives are expressed in terms of the space derivatives of moments of order 3. By a new application of the same process we can compute the time derivatives of these moments of order 3, but now they are defined in terms of the space derivatives of moments of order 4 and so on, and this goes on indefinitely. This chain reaction of moments of higher order on the evolution of moments of smaller order is known as the *closure problem* of turbulence theory (see Kraichnan [B.7.14]); it was already confronted by G. I. Taylor in 1935 [B.7.19]. The unavailability of a method of solution of this countable number of equations, simultaneously describing the evolution of all the moments, has suggested many approximate solutions. One of the most popular is the quasi-normality hypothesis from Millionstchikov [B.7.15]. The hypothesis is made that the moments of fourth order are in the same relation to the correlation as for a *normal* random vector field. This approximation has been systematically applied by Proudman and Reid [B.7.16] and Tatsumi [B.7.17].

The most heroic cure to cut the gordian knot is the introduction by E. Hopf [B.7.5] of the *characteristic functional*, getting rid, once and for all, of the moments; $\mathbf{v}^*(\mathbf{u})$ being an arbitrary linear functional of the velocity field $\mathbf{u}(\mathbf{x}, t, \omega)$, the characteristic functional is defined by

$$\Phi(\mathbf{v}^*) = E[\exp(i\mathbf{v}^*(\mathbf{u}))].$$

E. Hopf has given an integro-differential equation defining $\frac{\partial \Phi}{\partial t}$:

$$\frac{\partial \Phi}{\partial t} = L_2(\Phi) + \nu L_1(\Phi),$$

where L_1 is a linear differential operator (functional differentiation with respect to argument u_j of first order, which arises from viscosity terms in the Navier-Stokes equations) and L_2 is of second order and arises from the inertial terms.

Should it be possible to solve this equation, the knowledge of Φ being equivalent to the knowledge of the probability law of $\mathbf{u}(\mathbf{x}, t, \omega)$, this would yield a complete solution of the problem. But today one is far from this end and we agree with E. Hopf when he writes [B.7.7, p. 157] "The obstacles in the way of a rational theory of highly turbulent fluid flow are still formidable."

8. *Burgers Model.* The great mathematical difficulties encountered in the theory of turbulence seem insuperable today and at present the way to a successful attack on them seems hopelessly barred. However, there is no doubt, that many characteristic features of turbulent flows occur in a much larger class of similar nonlinear equations—"In order to gain insight into the nature of hydrodynamical phase flows we are, at present, forced to find and to treat simplified examples within this class" (E. Hopf [B.8.8, p. 304]). The most popular example is "Burgers model" [B.8.2] whose motion is defined by one scalar function $u(x, t)$ (in the half-plane $t > 0$, $-\infty < x < +\infty$) by the nonlinear equation:

$$u_t + uu_x = \nu u_{xx} \qquad (\nu = \text{constant} > 0). \tag{8.1}$$

This equation is the simplest analogue of the Navier-Stokes equations, insofar as it contains a nonlinear term with a space derivative and a linear term with a second-order space derivative multiplied by a factor which can be very small. Since 1940 extensive mathematical investigations have been carried out. From the beginning it was observed that the integrals exhibit the property that regions make their appearance where $|u_x|$ takes very high values (steep fronts). These can be considered to be analogous to the regions of very high vorticity in a turbulent flow.

The study of the integrals of the Burgers equation became easier when it was discovered (E. Hopf, 1946, J. D. Cole, 1949) that it can be reduced to the heat equation, by the transformation

$$u(x, t) = -2\nu \frac{\partial}{\partial x} \log w(x, t) \qquad w(x, t) > 0.$$

Starting from this fact, E. Hopf [B.8.9] proved an existence and

uniqueness theorem in the half-plane, even giving an explicit expression of $u(x, t)$, under the assumption that the initial condition $v(x)$ satisfies

$$\int_0^x v(\xi)\, d\xi = 0(x^2)$$

(which cannot be very much weakened as is shown by the example $u = x/t - \tau$, which explodes at $t = \tau > 0$).

From the explicit expression of $u(x, t)$, one easily deduces its asymptotic behavior for $t \to +\infty$, when $v(x) \in L(-\infty, +\infty)$. Most interesting is the consideration of the way in which the integral depends on the parameter ν and especially the behavior of $u(x, t, \nu)$ when $\nu \downarrow 0$. The connection between this limit and the integrals of $u_t + uu_{xx} = 0$ is illuminating. It throws great light on the coalescence of the steep fronts and the apparition of discontinuity lines.

If one wants to use the Burgers equation as a model for turbulence, the question naturally arises, are some integrals of (8.1) "complicated" enough to represent a chaotic flow? J. Bass [B.8.1] proved that $u(x, t)$ can be an almost periodic function of t and—what is more interesting—a "pseudo-random" (pseudo-aléatoire) function of t.

The Burgers equation being invariant under the translation $x \to x + h$, the consideration of the spatially homogeneous case is most natural. Starting from an initial stationary random function $v(x, \omega)$, the formula established by E. Hopf gives an explicit expression of the corresponding random integral $u(x, t, \omega)$. Unfortunately, no statistical property of $u(x, t, \omega)$ has yet been obtained in this rigorous way.

Burgers concentrates ([B.8.6, B.8.7]) his efforts on the moments

$$\overline{u_1^m u_2^n} = E[u(x, t, \omega)^m u(x + h, t, \omega)^n].$$

He finds the analog of von Kármán's equation for the propagation of correlation in homogeneous turbulence

$$\frac{\partial}{\partial t}(\overline{u_1 u_2}) = \frac{\partial}{\partial h}(\overline{u_1^2 u_2}) + 2\nu \frac{\partial^2}{\partial h^2}(\overline{u_1 u_2}); \tag{8.2}$$

a single equation to determine two functions! But here the situation is very different, because we know [from the explicit expression of $u(x, t, \omega)$] that it should be possible to determine $\overline{u_1 u_2}$ without additional hypotheses about $\overline{u_1^2 u_2}$. It is possible to obtain asymp-

totic expressions for these two moments. The most interesting fact is that the only statistical property of the random function $v(x, \omega)$ which appears in these expressions is the integral

$$J = \int_0^\infty \overline{u_1 u_2}\, dh,$$

which has been shown to be independent of t. In spite of the labor and ingenuity involved in these computations, it must be observed that the expressions obtained are asymptotic in two respects: the time t is very large and ν (which does not appear explicitly in the results) is very small.

Burgers [B.8.5] has also considered a different problem referring to the propagation along the x-axis of impulses introduced at $x = 0$ at a series of successive instants. In this case there is no homogeneity in space, but the system is statistically stationary with respect to time.

9. *Linear Equations; Well-Set Cauchy Problem.* The Navier-Stokes equations (and the simplified Burgers model) belong to the general class of partial differential equations

$$\frac{\partial u_j}{\partial t} = \mathfrak{A}_j[\mathbf{u}] \tag{9.1}$$

where the $\mathfrak{A}_j$ are differential operators in the space derivatives of the u_j; the coefficients do not depend explicitly on the time. This type of equation appears to be the most usual in the mechanics of continuous media. The building of a statistical mechanics of fluids is actually far beyond our reach because the $\mathfrak{A}_j$ are nonlinear. However, continuous media, whose motions are defined by *linear equations* have been the happy hunting ground of mathematical physics for a long time (vibrating string or beams, sound waves, Airy's gravity waves, and so on). Thus the question arises: is it possible to generalize the statistical mechanics of Gibbs to continuous media whose motions are defined by linear partial differential equations? We will sketch briefly the results obtained, referring to [B.9.4] for more details.

(a) The continuous medium fills a domain D_p in Euclidean space R^p: $\mathbf{x} = [x_1, \ldots, x_p]$; D_p has a smooth boundary B_{p-1}; one can have $D_p = R^p$.

(*b*) The state of the medium is defined, at any given time t, by a real-valued $\mathbf{u} = [u_1, \ldots, u_q]$ on D_p.

(c) For $0 < t < \tau \leq +\infty$, $\mathbf{u}(\mathbf{x}, t)$ satisfies the linear partial differential equations:

$$\frac{\partial u_j}{\partial t} = \sum_{k=1}^{q} P_{j,k}\left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_p}\right) u_k, \tag{9.2}$$

the $P_{j,k}$ being polynomials whose coefficients may depend on $\mathbf{x}$ but not on time t.

The determination of the motion corresponds to a typical Cauchy problem, when the initial conditions and the boundary conditions are given; the problem is *well set* (*bien posé*, J. Hadamard, 1903) if one can prove an *existence, uniqueness, and continuity theorem* for the integral $\mathbf{u}(\mathbf{x}, t)$. This proof can be achieved only when one has first defined

(α) the set $\mathfrak{U}_0$ of all functions $\mathbf{u}_0(\mathbf{x})$ on D_p considered as *admissible initial conditions;*

(β) the set $\mathfrak{U}_1$ of all functions $\mathbf{u}_1(\mathbf{x}, t)$ on $B_{p-1} \times (0, \tau)$ considered as *admissible boundary conditions;*

(γ) the set $\mathfrak{U}$ of all functions $\mathbf{u}(\mathbf{x}, t)$, taken as functions of $\mathbf{x}$ on D_p (for each t), considered as *regular integrals;*

(δ) topologies on the function spaces $\mathfrak{U}_0$, $\mathfrak{U}_1$, $\mathfrak{U}$:

(ϵ) the precise meaning of the limits

$$\mathbf{u}(\mathbf{x}, t) \to \mathbf{u}_1(\mathbf{x}, t) \text{ on } B_{p-1} \tag{9.3}$$

$$\mathbf{u}(\mathbf{x}, t) \to \mathbf{u}_0(\mathbf{x}) \text{ on } D_p \qquad \text{if } t \downarrow 0. \tag{9.4}$$

As a rule this choice is not unique. For instance, in the case of the heat equation in an infinite rod, in [B.7.13] three completely different choices are discussed, each one of them leading to a well-set Cauchy problem. This is of paramount importance because obviously the statistical mechanics will be strongly influenced by this initial step. We must not forget our liberty of choice, when later we are tempted to interpret our results as "laws of nature"; the "laws" are highly dependent on our initial fiat!

For simplicity we will here consider only the correspondence between $\mathfrak{U}_0$ and $\mathfrak{U}$; the boundary conditions appearing only in the background in most cases. For instance, they often reduce to $\mathbf{u}(\mathbf{x}, t) = 0$ on B_{p-1}, and they completely disappear if $D_p = R^p$.

Due to the linearity of the differential equation (9.2), we must obviously consider $\mathfrak{u}_0$ and $\mathfrak{u}$ as *linear topological spaces;* the correspondence between $\mathfrak{u}_0$ and $\mathfrak{u}_1$

$$\mathbf{u}(\mathbf{x}, t) = T_t \mathbf{u}_0(\mathbf{x}), \tag{9.5}$$

has the *semigroup property*

$$T_{t+s} = T_t T_s = T_s T_t. \tag{9.6}$$

This was noted as far back as 1924 by J. Hadamard and is a property which expresses the Huyghens principle ("principle of scientific determinism"; see [A.5.5, pp. 388–390]). The correspondence being continuous, by definition, T_t is a linear continuous transformation of $\mathfrak{u}_0$ unto $\mathfrak{u}$; hence the inverse image T_t^{-1} of any open set (in the topology $\mathfrak{I}$ on $\mathfrak{u}$) must be an open set in the topology $\mathfrak{I}_0$ on $\mathfrak{u}_0$, for any $t \in (0, \tau)$.

Most often $\mathfrak{u}$ is a proper subset of $\mathfrak{u}_0$, the regular integrals being smoother than the initial conditions; thus for the heat equation in an infinite rod if $\mathfrak{u}_0 = C[-\infty, +\infty,]$ or $L^p[-\infty + \infty]$, then $\mathbf{u}(\mathbf{x}, t)$ belongs, for each $t > 0$, to the same space, but must be an entire function of $\mathbf{x}$.

Thus the function space $\mathfrak{u}_0$ can be taken as phase space Ω. From each point ω of Ω there issues one and only one trajectory representing the evolution of the continuous medium. Here we have a very natural generalization of the description of motion of a holonomic system. Nevertheless we must point to an important difference; the Hamilton-Jacobi equations (1.1) are invariant if t is replaced by $-t$ (the past and the future are both determined by the present). Such is not the case for all partial differential equations (9.2); for instance, it is not true for equations of parabolic type. This explains why the group of transformations (1.2) are replaced, in general, by a semigroup.

The next step is to define *probability* measures on Ω. The most natural way seems to be to define a *regular* probability measure P (see page 296) on the linear topological space $\mathfrak{u}_0$; hence, putting

$$P_t(\Theta) = P[T_t^{-1}\Theta] \tag{9.7}$$

for all open sets of $\mathfrak{u}$, we construct a σ-additive measure on the Borel sets of $\mathfrak{u}$. By definition, the Lebesgue completion of this will be a regular measure in the space of regular integrals $\mathfrak{u} \subset \mathfrak{u}_0$

[B.9.5]. We obtain the random integrals of (9.2)

$$\mathbf{u} = \mathbf{u}(\mathbf{x}, t, \omega),$$

if we assume that ω, which represents the initial conditions $\mathbf{u}_0(\mathbf{x})$ in $\Omega = \mathfrak{U}_0$, is chosen at random according to the probability law P.

In order that the continuous medium be in *statistical equilibrium* or, what is the same, that the random integral be a *stationary random function* of t, we must have

$$P_t(\mathcal{O}) = P(\mathcal{O})$$

for all open sets. We will see in an example that for some parabolic equations this is not possible, except for trivial measures, giving the measure one to the random constants $\mathbf{u}(\omega)$ independent of $\mathbf{x}$ and t.

10. *Banach Spaces; Abstract Cauchy Problem.* We will now *specialize* to linear partial differential equations (9.2) *with constant coefficients.*

For a great number of well-set Cauchy problems, corresponding to systems of linear equations with constant coefficients one is most naturally led for the phase space Ω to a *Banach space;* the most usual spaces being $C^{(m)}(D_p)$ and $L^m(D_p)$.

Starting from this remark E. Hille gave a new formulation of the problem as an *abstract Cauchy problem* [B.10.3]. For each t, $\mathbf{u}(\mathbf{x}, t)$ is an element ω of the Banach space Ω and the integral corresponding to $\mathbf{u}_0(\mathbf{x})$ is defined by the trajectory originating at $\mathbf{u}_0$ under the action of a suitable semigroup of continuous linear transformations T_t of Ω into itself. In [A.5.5] E. Hille developed important applications to hyperbolic, parabolic, and elliptic equations with constant coefficients. R. S. Phillips [B.10.7] modified the formulation somewhat and in [A.5.6, p. 619] an abstract Cauchy problem is described as follows: Given a linear operator $\mathfrak{A}(\omega)$, with domain $\mathfrak{D}(\mathfrak{A})$ and range $\mathfrak{R}(\mathfrak{A})$ in the Banach space Ω, and given $\omega_0 \in \Omega$, find a function $\omega(t)$ from $(0, +\infty)$ to Ω such that:

(a) $\omega(t)$ is strongly absolutely continuous and continuously differentiable in each finite subinterval of $(0, +\infty)$;

(b) for each $t > 0$, $\omega(t) \in \mathfrak{D}(\mathfrak{A})$ and

$$\frac{d\omega}{dt} = \mathfrak{A}[\omega(t)]; \tag{10.1}$$

(c) $$\lim_{t \downarrow 0} \|\omega(t) - \omega_0\| = 0.$$

We refer to [A.5.6, pp. 619–633] for existence and uniqueness theorems, and for the properties of the semigroup of transformations T_t [the so-called C_0-type, according to condition (c)]. The differential operator $\mathfrak{A}$ is the infinitesimal generator of a strongly continuous semigroup of transformations T_t of the C_0-type if

(10.2) $$\lim_{h \downarrow 0} \left\| \frac{1}{h} (T_h\omega - \omega) - \mathfrak{A}(\omega) \right\| = 0.$$

The problem of integration of the linear equations (9.2) being solved as an abstract Cauchy problem, if we define a probability measure P on the Banach space Ω, we obtain an elegant representation of the random integrals as random functions $\omega(t)$ from $[0, +\infty]$ to Ω. Most of the observable *properties* of a continuous medium being defined by linear functionals, we will suppose that P is an L-measure (M. Fréchet, E. Mourier), that is, that every linear functional is P-measurable; the connections of the L-measures with the regular measures are given by [B.9.5, p. 667]:

(a) A measure on a Banach space is an L-measure if, and only if, it is an extension of a regular measure for the weak topology.

(b) A measure on a separable Banach space is an L-measure if, and only if, it is an extension of a regular measure in the strong topology.

In all problems of interest in theoretical physics the construction of an L-measure on the Banach space is greatly simplified by the existence of a *Schauder basis* $[e_1, \ldots, e_n, \ldots]$: there is a one-one correspondence S between the point ω of Ω and a sequence of real numbers $[\eta_1, \ldots, \eta_n, \ldots] \subset R^\infty$ such that:

(10.3) $$\lim_{n \to \infty} \|\omega - (\eta_1 e_1 + \eta_2 e_2 \cdots + \eta_n e_n)\| = 0.$$

We start from a sequence of random variables defined by the sequence of probability distributions

$$\text{Prob}\,[\eta_1 < \alpha_1, \ldots, \eta_n < \alpha_n] = \mathcal{F}_n(\alpha_1, \ldots, \alpha_n).$$

By a classical theorem (Kolmogorov) this determines a unique probability measure ν on a σ-algebra $\mathcal{G}$ of subsets of R^∞; then Ω being, through (10.3), in a one-one correspondence with a subset

Ω^* of R^∞, $\Omega^* = S\Omega$, we give a condition necessary and sufficient for

$$\Omega^* \in \mathcal{G} \qquad \text{and } \nu(\Omega^*) = 1.$$

The probability measure P is now defined on the σ-algebra $\mathcal{F}$ of subsets A of Ω such that $SA \in \mathcal{G}^*$ ($\mathcal{G}^*$ = restriction of $\mathcal{G}$ to Ω^*) by

$$P(A) = \nu(SA)$$

and finally, we prove that P is an L-measure (for details we refer to [B.10.4]).

When, for a continuous medium, the phase space Ω is a Banach space with a Schauder basis, the state of the system is thus defined by a *countable number* of coordinates $[\eta_1, \ldots, \eta_n, \ldots]$; in this case one has the most natural extension of Gibbs' statistical mechanics.

Among the L-measures on a Banach space, the *normal* measures are particularly important; these are the measures for which every linear functional is a normal random variable (M. Fréchet). As for the canonical distribution of Gibbs, the choice of a normal measure for the statistical mechanics of a continuous medium is justified by the principle of maximum information entropy (see [B.10.5] where we give a proof for the vibrating string).

As a last remark let us point out that in many cases the solution of the abstract Cauchy problem becomes much easier if one uses the Fourier-transform technique (which amounts to the replacement of the Banach space Ω by another isomorphic Banach space). Púrsuing the ideas of E. Hille [B.10.3], G. Birkhoff, and T. W. Mullikin [B.10.1] have shown how to construct a C_0-semigroup for any well set Cauchy problem for linear partial differential equations with constant coefficients. In [B.10.2] G. Birkhoff makes, by means of the Fourier-transform technique, a thorough study of the case in which D_p is the product of s real lines and $p - s$ circles, which represents the most general locally Euclidean Abelian group manifold.

11. *Spatial Homogeneity.* The linear partial differential equations (9.2), with constant coefficients, are invariant for any translation $\mathbf{x} \to \mathbf{x} + \mathbf{h}$ in R^p. Thus, when $D_p = R^p$, the class of random integrals which comes to mind is obviously the class of $\mathbf{u}(\mathbf{x}, t, \omega)$ which are stationary with respect to $\mathbf{x}$ at each instant t.

A regular measure P on a function space, whose topology is invariant under the translations $\mathbf{x} \to \mathbf{x} + \mathbf{h}$, will be called *homogeneous* if $P(B + \mathbf{h}) = P(B)$ for any Borel set B.

One can prove [B.9.5] the following result: if P is a homogeneous regular measure on the linear topological space $\mathfrak{U}_0$ (initial conditions) and if P_t is defined (for each $t > 0$) on the linear topological space $\mathfrak{U}$ (regular integrals) by the condition (9.5), then P_t is a homogeneous regular measure for each $t > 0$. In other words, if the initial conditions $\omega = \mathbf{u}_0(\mathbf{x})$ are spacially homogeneous, then the random integrals $\mathbf{u}(\mathbf{x}, t, \omega)$ are stationary with respect to $\mathbf{x}$ for any $t > 0$.

If the equations (9.2) are interpreted as defining the motion of a continuous medium, the consideration of the total energy of the system is of paramount importance. It is only when this total energy is finite that it is natural to choose for the phase space a Banach space and to interpret the problem as an abstract Cauchy problem. When the total energy is infinite, the method described in Section 10 cannot be applied.

The simplest case occurs for continuous media such that the energy corresponding to a cube K has the same expression as the kinetic energy of an incompressible fluid.

$$E_K = \int_K |\mathbf{u}(\mathbf{x})|^2 \, dm(\mathbf{x}).$$

Due to spatial homogeneity, the cubes K and $K + \mathbf{h}$ have, on the average, the same energy. Thus, the mean total energy for the whole space is infinite and we are prohibited from taking as phase space the Hilbert space $L^2(R^p)$.

It is most natural to define the topology by the countable family of seminorms

$$\|\mathbf{u}\|_n = \left[\int_{K_n} |\mathbf{u}(\mathbf{x})|^2 \, dm(\mathbf{x})\right]^{1/2} \tag{11.1}$$

where the K_n, $n = 1, 2, \ldots$ are the cubes $\{\mathbf{x}: |x_j| \leq n, j = 1, 2 \ldots, p\}$. This linear topological space is precisely the space Λ, introduced on page 296. Alternately, one can also define Λ as the unrestricted direct topological product of countable many Hilbert spaces $L^2(\Delta_n)$; $\{\Delta_1, \ldots, \Delta_n \ldots\}$ being a countable partition of R^p into compact domains [B.7.3 p. 668]. This observation indicates that, even though Λ is not a Banach space, a point of Λ can nonetheless be defined by a countable number of coordinates. Thus, exactly as in Section 10, the continuous media having Λ as phase space lead to the most simple generalization of the statistical

mechanics of Gibbs to systems with a countable number of degrees of freedom.

Finally, let us note that Λ is complete, locally convex, and metrizable (separable Fréchet space).

Now let us sketch briefly the simplest instance of stationary random integrals in Λ. For $p = q = 1$, the equations (9.2) reduce to

$$\frac{\partial u}{\partial t} = Q\left(\frac{\partial}{\partial x}\right) u. \tag{11.2}$$

We suppose that the polynomial Q satisfies the condition

$$\lim_{|\kappa| \to \infty} Q(i\kappa) = -\infty \qquad \kappa \text{ real},$$

which means that the linear differential equation is of the *parabolic type*. Now we shall define a probability measure P on Λ. Let us take an arbitrary absolutely continuous energy spectrum

$$d\mathcal{E}(\kappa) = \mathcal{E}'(\kappa)\, d\kappa \qquad \mathcal{E}' > 0 \qquad \text{and } \mathcal{E}' \in L(0, +\infty). \tag{11.3}$$

From the uniqueness theorem [B.7.3, p. 698], we know that $\mathcal{E}(\kappa)$ determines a unique *normal homogeneous admissible regular measure* P. Let us assume that the sample functions $\omega = \omega(x)$ (the initial conditions) are chosen at random in Λ according to

$$\text{Prob}\,[\omega \in A] = P(A).$$

The function

$$G(x, t) = \int_{-\infty}^{+\infty} e^{i\kappa x + tQ(i\kappa)}\, d\kappa \qquad t > 0$$

being the fundamental integral or the Green's function of (11.2), *the formula*

$$u(x, t, \omega) = \int_{-\infty}^{+\infty} G(x - \xi, t)\omega(\xi)\, d\xi \qquad t > 0 \tag{11.4}$$

then defines a random integral of (11.2) *for almost all* ω. It is an analytic function of (x, t) and its derivatives satisfy (11.2) (literal or smooth integral according to the current nomenclature).

The random integral $u(x, t, \omega)$ is stationary with respect to x for each $t > 0$, and it has an absolutely continuous spectrum

$$d\mathcal{E}_u(\kappa) = \mathcal{E}'(\kappa) \exp\,[2tReQ(i\kappa)]\, d\kappa.$$

Here we have an example of a *statistically well-set Cauchy problem.*

The space Λ is an enormous bag containing far too many functions, but by using our statistics as a sieve, we discard many of them. The initial function being submitted only to the condition $\omega(x) \in L^2[a, b]$ for any finite $[a, b]$, one can show that (11.4) has no meaning in general; but among all the functions $\omega(x) \in \Lambda$ we select a particular set Λ^*, to which our statistics gives measure one

$$\Lambda^* = \left\{\omega : \frac{\omega(x)}{\sqrt{1+x^2}} \in L^2(-\infty, +\infty)\right\},$$

and for these special initial conditions the integral (11.4) gives the solution of the Cauchy problem. All the other functions are unimportant statistically because $P(\Lambda - \Lambda^*) = 0$. The special set Λ^* is singled out through *metric transitivity*, which is realized here as a consequence of the *absolute continuity* of the *spectrum* $\mathcal{E}(\kappa)$ [B.7.3, p. 331]. Due to this metric transitivity,

$$\lim_{A\to\infty} \frac{1}{2A} \int_{-A}^{A} \omega(x)^2\, dx = \mathcal{E}(+\infty) < +\infty$$

for almost all ω. But it is known [C.4, p. 138] that if, for a function $f(x)$ the integral

$$\frac{1}{2A} \int_{-A}^{A} f(x)^2\, dx$$

is bounded for $A \to \infty$, then $f(x)/\sqrt{1+x^2} \in L^2(-\infty, +\infty)$. Thus for any normal homogeneous regular measure P on Λ one has

$$P(\Lambda^*) = 1.$$

Now a question arises which is of great importance for statistical mechanics. The random integral (11.4) is stationary with respect to x. May we choose the measure P so that $u(x, t, \omega)$ will also be stationary with respect to t? The answer is: no, except in trivial situations. For instance, for the heat equation [B.9.4, p. 180], $u(x, t, \omega)$ can be stationary with respect to x and t only if P gives measure one to the set of constant initial functions $\omega(x) = a$. This result points to a major difference between parabolic and hyperbolic equations. For instance, there are large classes of random integrals of the wave equations which are at the same time stationary with respect to x and t (for example, sound waves [B.9.4, p. 193]). For an ensemble of systems whose motion is defined by

a hyperbolic equation, it is always possible to realize statistical equilibrium. This is not true for systems corresponding to a parabolic equation, except in the trivial case when each system of the statistical ensemble is itself in equilibrium. This remark could have far-reaching consequences in physics.

12. *Conclusions.* The reader has certainly observed that, in the preceding sections, we have always looked through a window opening on a very limited horizon. From the beginning we have been, so to speak, polarized by statistical mechanics. We must now stress that there are many other ways leading to random integrals of partial differential equations, and most of these ways lie on virgin soil.

For instance, let us take the random integrals of Laplace's equation in the unit circle.

(1) One can take the Dirichlet problem: The harmonic function in the circle $r < 1$ is determined by its boundary values $\omega(\theta)$ $0 \leq \theta < 2\pi$ through the Poisson or Poisson-Stieltjes integral. Here one is almost in the same situation as in the Cauchy problem of the preceding sections. The boundary condition plays the same role as the initial condition; a probability measure on the set $\mathfrak{U}_1$ of the boundary condition induces a measure on the set of harmonic functions [B.11.3].

(2) The harmonic function being defined by the series

$$\tfrac{1}{2}a_0 + \sum_{n=1}^{+\infty} r^n(a_n \cos n\theta + b_n \sin n\theta),$$

one assumes that the coefficients are random variables $a_n(\omega)$, $b_n(\omega)$ such that

$$\text{Prob}\,[\limsup |a_n|^{1/n} = 1] = 1 \quad \text{and} \quad \text{Prob}\,[\limsup |b_n|^{1/n} = 1] = 1.$$

This is a completely different way of defining a random harmonic function in the circle $r < 1$.

(3) Several different types of Banach spaces of harmonic functions have been considered (see [B.12.4]). On one of them, $\mathfrak{B}$, define an L-measure. The random element u in $\mathfrak{B}$ corresponding to this measure constitutes a new type of random integral of the Laplace equation.

(4) One can replace the Banach space by a linear topological

space analogous to the space Λ. On the set of harmonic functions in the unit circle γ, define a topology by the countable family of seminorms

$$\|u\|_n = \left[\int_{\gamma_n} |u(r, \theta)|^2 r\, dr\, d\theta\right]^{1/2},$$

$[\gamma_1, \ldots, \gamma_n, \ldots]$ being an increasing sequence of circles

$$\gamma_n \subset \gamma_{n+1}, \qquad \gamma_n \uparrow \gamma.$$

We do not know of any research exploring the two last fields.

I shall not endeavor to discuss the connections between the random integrals considered in this survey and the research made, more or less, in the spirit of the Monte-Carlo method. In the latter case one explores integrals of partial differential equations (for instance harmonic functions) by means of a random walk or of Brownian motion path. The papers [B.12.1, B.12.2, and B.12.3] by J. L. Doob are among the most typical of this point of view. Their results, for instance in an abstract theory of boundary value problems, possibly have certain connections with some of the ideas related here; but the task of analyzing them would prove overwhelming for the time alloted to this survey and for my own limitations.

C. APPENDIX

I. In the usual terminology of measure theory, a *probability space* is a measure space such that $P(\Omega) = 1$.

A *random variable* $x(\omega)$ is a P-measurable function from Ω to a space $\mathfrak{X}$. Here we take for $\mathfrak{X}$ the real line R and exceptionally the complex plane; the only exception is Section 9, where $\mathfrak{X}$ is a Banach space.

Given a set J, a *random function* on J is a set of random variables $x(t, \omega)$ defined for all $t \in J$. Here one always has $J \subset R^p$. In this note we will develop mostly the case $J \subset R$. For a fixed ω the function $x(t, \omega)$ from J to R is a *sample* (or a realization) of the random function.

Taking for Ω a given set of functions $\omega(t)$, $t \in J$, defining a probability measure P on Ω, one obtains a random function $x(t, \omega)$ whose samples are the given functions $\omega(t)$. One says that $x(t, \omega)$ is of the *function space type*.

The *probability law* of $x(t, \omega)$ is defined by the set of functions

$$\text{(C.1)} \quad \mathfrak{F}_n(\xi_1, \ldots, \xi_n, t_1, \ldots, t_n) = P\{\omega : x(t_1, \omega) < \xi_1, \ldots x(t_n, \omega) < \xi_n\}$$

$n = 1, 2, \ldots$, $\{t_1, \ldots, t_n\} \subset J$, $\{\xi_1, \ldots, \xi_n\} \subset R$ arbitrary.

The *expectation* of a random variable is, by definition,

$$E[x(\omega)] = \int_\Omega x(\omega)\, dP \qquad \text{if } x \in L(\Omega).$$

If $x(t, \omega) \in L^2(\Omega)$ for $t \in J$, the *mean* and *covariance* are defined by

$$\text{(C.2)} \qquad m(t) = E[x(t, \omega)]$$

$$\text{(C.3)} \qquad \Gamma(t, s) = E[[x(t, \omega) - m(t)][x(s, \omega) - m(s)]].$$

A random function is *normal* if its probability law (C.1) is normal (= Gaussian). The mean and variance completely determine the probability law.

If one knows only the probability law of a random function, one finds that the probabilities that the samples are bounded in an interval, or continuous, or derivable, etc., may not be defined because the corresponding ω set may not be measurable; to cope with this major difficulty the concept of *separability* has been introduced (J. L. Doob). Separability means roughly that the samples are as regular as their restriction to some countable dense subset of J. The random function $x(t, \omega)$ is called *separable* if there exists a sequence $\{t_1, \ldots, t_n, \ldots\}$ dense in J and a set $N \subset \Omega$ of probability 0, such that for any closed interval $[a, b] \subset R$ and any open interval $(\alpha, \beta) \subset J$ one has

$$\{\omega : x(t_j, \omega) \in [a, b], \forall t_j \in (\alpha, \beta)\} - \{\omega : x(t, \omega) \in [a, b], \forall t \in (\alpha, \beta)\} \subset N.$$

The great interest of this definition arises from the proposition: Given any random function $x(t, \omega)$ there exists a separable random function $\bar{x}(t, \omega)$ such that for each fixed t:

$$P\{\omega : \bar{x}(t, \omega) = x(t, \omega)\} = 1.$$

$\bar{x}(t, \omega)$ is called the *separable version* of $x(t, \omega)$; obviously $\bar{x}(t, \omega)$ has the same probability law as $x(t, \omega)$. When we say that a given random function (for example, the integral of a differential equation) is continuous, this is always meant implicitly for the separable version.

A random function is *measurable* if $x(t, \omega)$ is measurable with respect to the product measure $m \times P$(m = Lebesgue measure on J).

The condition

$$\text{(C.4)} \qquad \lim_{|t-s|\to 0} [\Gamma(t, t) + \Gamma(s, s) - 2\Gamma(t, s)] = 0 \qquad t \in J \; s \in J$$

implies that the separable version $x(t, \omega)$ is measurable. If $x(t, \omega)$ is measurable and if $\Gamma(t, t) \in L(J)$, then almost all samples belong to $L^2(J)$.

II. A random function is *stationary* if its probability law is invariant for any translation

$$\mathfrak{F}_n(\xi_1, \ldots, \xi_n, t_1 + h, \ldots, t_n + h) = \mathfrak{F}_n(\xi_1, \ldots, \xi_n, t_1, \ldots, t_n)$$

for all h.

If $x(t, \omega) \in L^2(\Omega)$, then

$$\Gamma(t, s) = R(t - s);$$

R is called the *correlation*. The continuity of $R(h)$ for $h = 0$ implies its uniform continuity in any finite interval and one has

$$R(h) = \int_0^\infty \cos \kappa h \, d\mathcal{E}(\kappa).$$

The nondecreasing and bounded function $\mathcal{E}(\kappa)$ is called the *energy spectrum* of $x(t, \omega)$.

For a *normal* random function, the absolute continuity of $\mathcal{E}(\kappa)$ implies *metric transitivity* and one has:

$$\lim_{A\to+\infty} \frac{1}{2A} \int_{-A}^{A} x(t, \omega)x(t + h, \omega) \, dt = R(h)$$

for almost all samples.

III. A random function is of the *Markov type* (a *Markov process*) if the conditional probability

$$P\{\omega : x(t, \omega) < \eta | x(s_1, \omega) = \xi_1, \ldots, x(s_n, \omega) = \xi_n, x(s, \omega) = \xi\}$$

(for any $s_1 < s_2 \cdots < s_n < s < t$) is equal to the conditional probability

$$P\{\omega : x(t, \omega) < \eta | x(s, \omega) = \xi\} = P(s, \xi, t, \eta) \qquad s < t.$$

When the present $x(s, \omega)$ is known, the knowledge of the past $x(s_1, \omega), \ldots, x(s_n, \omega)$ does not add any information on the future $x(t, \omega)$. The function $P(s, \xi, t, \eta)$ is called the *transition probability*, and the knowledge of the transition probability and of the initial probability $P\{\omega : x(0, \omega) < \zeta\} = P_0(\zeta)$ completely determines the probability law of the Markov process in $[0, +\infty)$.

The Huyghens principle for a Markov process is expressed by the *Chapman-Kolmogorov equation*

$$P(s, \xi, t, \eta) = \int_{-\infty}^{+\infty} P(\tau, \zeta, t, \eta)\, d_\zeta P(s, \xi, \tau, \zeta) \qquad s < \tau < t.$$

IV. One says that a random function has *independent increments* if for any set of nonoverlapping intervals $[s_1, t_1], \ldots (s_n, t_n]$

$$s_1 < t_1 \leq s_2 < t_2 \leq \cdots t_{n-1} \leq s_n < t_n$$

the increments.

$$x(t_1, \omega) - x(s_1, \omega), \ldots x(t_n, \omega) - x(s_n, \omega)$$

are independent random variables. The *Wiener-Lévy function* or *Brownian process* is one of the most important examples. The random function $W(t, \omega)$ is defined by the conditions

(a) $W(0, \omega) = 0$
(b) $W(t, \omega)$ has independent increments
(c) $W(t, \omega) - W(s, \omega)$ follows the normal law with

$$E[W(t, \omega) - W(s, \omega)] = 0$$

$$E[(W(t, \omega) - W(s, \omega))^2] = t - s \qquad 0 \leq s < t.$$

One deduces from these conditions that $W(t, \omega)$ is a *normal random function* and that

$$m(t) = E[W(t, \omega)] = 0$$

$$\Gamma(t, s) = E[W(t, \omega)W(s, \omega)] = \text{Inf } (t, s);$$

moreover, $W(t, \omega)$ is a Markov process.

Almost all samples of $W(t, \omega)$ are continuous in any finite interval, but are not of bounded variation; $W(t, \omega)$ has no derivative; more precisely each sample has a finite derivative for at most a set of t of measure zero. One has, for almost all ω,

$$\limsup_{h \downarrow 0} \frac{|W(t + h, \omega) - W(t, \omega)|}{\sqrt{2h \log|\log h|}} = 1.$$

In [C.1] (1957) we gave a representation of $W(t, \omega)$ in $[0, 1]$ by a series of triangular functions $e_n(t)$ [Schaudèr basis for $C[0, 1]$] which converges almost surely absolutely and uniformly in t.

$$W(t, \omega) = \Sigma_0^{+\infty} \eta_n(\omega) e_n(t) \tag{C.5}$$

where the $\eta_n(\omega)$ are independent normal random variables:

$$E(\eta_n) = 0,\ E(\eta_0^2) = \frac{1}{2},\ E(\eta_n^2) = \frac{1}{2^{q+1}} \qquad 2^{q-1} \leq n \leq 2^q - 1.$$

From this representation it is easy to develop an elementary theory of the Wiener-Lévy function [C.2] and in particular [C.3] to give a definition of the pseudo-Stieltjes integrals (W is not of bounded variation!)

$$\int_0^1 F(t)\, dW(t, \omega) \qquad \text{and} \qquad \int_0^{+\infty} F(t)\, dW(t, \omega)$$

by an almost surely convergent series;
for instance if $F(t) \in L^2[0, 1]$ we define the integral

$$X(\omega) = \int_0^1 F(t)\, dW(t, \omega)$$

as the sum of the series:

$$X(\omega) = \Sigma_{n=0}^{n=+\infty} 2^{(q+1)/2} a_n \eta_n(\omega)$$

where the a_n are the Fourier coefficients of $F(t)$ with respect to the Haar functions $h_n(t)$ (orthonormal basis in $L^2[0, 1]$, connected with the Schauder basis in $C[0, 1]$ by: $e_n(t) = 2^{(q_n+1)/2} \int_0^t h_n(s)\, ds$); the series is almost surely convergent and one has:

$$E(X) = 0 \qquad E(X^2) = \Sigma_0^{+\infty} a_n{}^2 = \int_0^1 F(t)^2\, dt.$$

REFERENCES

G. *General References*

G.1. *Probability*

G.1.1. J. L. Doob, *Stochastic Processes*, 1st ed., New York (1953) Wiley.
G.1.2. E. B. Dynkin, *Theory of Markov processes*, London (1960) Pergamon Press.
G.1.3. P. Lévy, *Processus stochastiques et mouvement brownien*, Paris (1948) Gauthier Villars.

G.1.4. M. Loève, *Probability theory*, 3rd ed., Princeton, N.J. (1963) Van Nostrand.

G.1.5. J. Neveu, *Bases mathématiques du Calcul des Probabilités*, Paris (1964) Masson.

G.2. *Differential Equations*

G.2.1. G. Birkhoff and G. C. Rota, *Ordinary differential equations*, 3rd ed., Boston (1962) Ginn.

G.2.2. A. Friedman, *Generalized functions and partial differential equations*, Englewood Cliffs, N.J. (1963) Prentice-Hall.

G.2.3. L. Hormander, *Linear partial differential operators*, Berlin (1963) Springer Verlag.

G.3. *Statistical Mechanics*

G.3.1. A. Blanc-Lapierre, P. Casal, and A. Tortrat, *Méthodes mathématiques de la mécanique statistique*, Paris (1959) Masson.

G.3.2. A. I. Khinchin, *Mathematical foundations of statistical mechanics*, New York (1949) Dover.

A. *Ordinary Differential Equations*

A.1. *Statistical Mechanics*

A.1.1. J. W. Gibbs, *Elementary principles in statistical mechanics*, New Haven (1902) Yale Univ. Press.

A.1.2. J. Kampé de Fériet, *A lecture course on information theory and statistical mechanics*, Dept. of Applied Math., Indian Institute of Science, Bangalore (1963).

A.2. *Ordinary Differential Equations*

A.2.1. G. K. Batchelor, Diffusion in a field of homogeneous turbulence, I. Eulerian analysis, *Austral. J. Sci. Res.*, Ser. A, *Phys. Sci.*, **2** (1949) 437–450.

A.2.2. G. K. Batchelor, Diffusion in a field of homogeneous turbulence, II. The relative motion of particles, *Proc. Cambridge Phil. Soc.*, **48** (1962) 345, 362.

A.2.3. Yu. K. Belyaev, Local properties of a stationary gaussian process, *Teoriya Veroyatnostei*, **5** (1960) 117–120.

A.2.4. Yu. K. Belyaev, On the continuity and differentiability of realizations of gaussian processes, *Teoriya Veroyatnostei*, **6** (1961) 340–342.

A.2.5. Yu. K. Belyaev, Properties of the sample function of a stationary random process, *Teoriya Veroyatnostei*, **5** (1960) 120–122.

A.2.6. A. T. Bharucha-Reid, On random solutions of integral equations in Banach spaces, *Trans. 2nd Prague Conf. Inform. Theory, Statistical decision functions random processes* New York (1961) Academic press, 37–46.

A.2.7. A. T. Bharucha-Reid, On random operator equations in Banach spaces, *Bull. Acad. Polon. Sci., Ser. Sci. Math. Astr. Phys.*, **7** (1959) 561–564.

A.2.8. A. T. Bharucha-Reid, Sur les équations intégrales aléatoires de

Fredholm à noyaux séparables, *C. R. Acad. Sci. Paris*, **250** (1960) 454–456, 657–658.

A.2.9. R. H. Cameron, Differential equations involving a parametric function, *Proc. Amer. Math. Soc.*, **8** (1957) 834–840.

A.2.10. R. H. Cameron, Nonlinear Volterra functional equations and linear parabolic differential systems, *J. d'Anal. Math.*, **5** (1957) 136–182.

A.2.11. J. Delporte, Dérivabilité presque sûre des fonctions aléatoires normales (unpublished paper).

A.2.12. G. Hunt, Random Fourier transforms, *Trans. Amer. Math. Soc.*, **71** (1951) 38–69.

A.2.13. K. Karhunen, Lineare transformationen stationärer stochasticher Prozesse, 10*e congrès mathémat scandinaves, Copenhague* (1946) 320–324.

A.2.14. N. N. Krylov and N. N. Bogolyubov, Concerning a Fokker-Planck equation occurring in perturbation theory treated by a method based on the spectral properties of the perturbed Hamiltonian (Russian), *Zap. Kafedny Mat. Fiz. Akad. Nauk. Ukrain. R.S.R.*, **4** (1939) 5–158.

A.2.15. G. Ya Lubarskii and Yu. L. Robotnikov, On the theory of differential equations with random coefficients, *Teoriya Veroyatnostei*, **8** (1963) 290–298.

A.2.16. J. C. Samuels and A. C. Eringen, On stochastic linear systems, *J. Math. and Phys.*, **38**, 2 (1959) 83–104.

A.3. *Stochastic Differential Equations*

A.3.1. L. Bachelier, Théorie de la spéculation, *Ann. Sci. Ecole Norm. Sup.*, **3** (1900) 21–86.

A.3.2. S. Bernstein, *C. R. Acad. Sc. U.R.S.S., N.S.* (1934) 1–9, 361–364.

A.3.3. S. Bernstein, Equations différentielles stochastiques, *Act. Sci. Ind.*, **738** (1938) 5–31.

A.3.4. J. L. Doob, The brownian motion and stochastic equations, *Ann. Math.*, **43** (1942), 351–369.

A.3.5. K. Ito, Stochastic integral, *Proc. Imp. Acad. Tokyo*, **20** (1944) 519–524.

A.3.6. K. Ito, On a stochastic integral equation, *Proc. Jap. Acad.*, **1–4** (1946) 32–35.

A.3.7. K. Ito, On stochastic differential equations, *Mem. Amer. Math. Soc.*, **4** (1951).

A.3.8. M. Kac, Random walk and the theory of the brownian motion, *Amer. Math. Monthly*, **54**, 295–317.

A.3.9. P. Lévy, *Le mouvement brownien, Mém. Sc. Math.*, **126** (1954).

A.3.10. P. Lévy, Le mouvement brownien fonction d'un ou de plusieurs paramètres, *Rend. Matematica*, **22**, (1963) 24–101.

A.3.11. G. Maruyama, Continuous Markov processes and stochastic equations, *Rend. Circ. Math. Palermo*, **4** (1955) 48–90.

A.3.12. R. C. Paley and N. Wiener, *Fourier transforms in the complex domain, Amer. Math. Soc. Coll.*, **19** (1934).

A.3.13. A. V. Sokorod, On the existence and uniqueness of solutions of stochastic differential equations, *Sibvisk. Mat. Z.*, **2** (1961) 129–137.

A.3.14. G. E. Uhlenbeck and L. S. Ornstein, On the theory of the brownian motion, *Phys. Rev.*, **36** (1930) 823–841.

A.3.15. N. Wiener, Differential spaces, *J. Mass. Phys. Mass. Inst. Techn.*, **2** (1923) 131–174.

A.3.16. N. Wiener, Generalized harmonic analysis, *Acta. Math.*, **55** (1930) 117–258.

A.4. *Transition Probabilities and Diffusion Equations*

A.4.1. W. Feller, Zur theorie des Stochasticher Prozesse (Existenz und Eindeutigkeitsätze), *Math. Ann.*, **113** (1936) 113–160.

A.4.2. R. Fortet, Les fonctions aléatoires du type de Markoff associées à certaines équations linéaires aux dérivées partielles du type parabolique, *J. Math. pur. et appl.*, 9*e série*, **22** (1943) 117–243.

A.4.3. A. Fuchs, Continuité des fonctions aléatoires de Markov, *Bull. de la soc. math. France*, **88** (1960) 157–216.

A.4.4. A. N. Khintchine, Asymptotische Gesetze der Wahrscheinlichkeitsrechnung, *Ergeb. Math.*, **4** (1933).

A.4.5. A. N. Kolmogoroff, Analytische methodes der Wahrscheinlichkeitsrechnung, *Math. Ann.*, **104** (1931) 415–458.

A.4.6. Ming Chen Wang and G. E. Uhlenbeck, On the theory of the brownian motion, *Rev. Mod. Phys.*, **17** (1945) 323–342.

A.5. *Semigroup Theory*

A.5.1. E. B. Dynkin, Markov processes and semi-groups of operators, *Teoriya Veroyatnostei*, **1** (1956) 22–33.

A.5.2. E. B. Dynkin, Infinitestimal operators of Markov processes, *Teoriya Veroyatnostei*, **1** (1956) 34–54.

A.5.3. W. Feller, The general diffusion operator and positivity preserving semi-groups in one dimension, *Ann. Math.*, **60** (1954) 417–436.

A.5.4. W. Feller, On second order differential equations, *Ann. Math.*, **61** (1955) 90–105.

A.5.5. E. Hille, *Functional analysis and semi-groups, Amer. Math. Soc. Coll.*, **31** (1948).

A.5.6. E. Hille and R. S. Phillips, *Functional analysis and semi-groups, Amer. Math. Soc. Coll.*, **31** (1957).

A.5.7. A. Neveu, Théorie des semi-groupes de Markoff, *Univ. Calif. Public. in Statist.*, **2** (July 1958) 319–394.

A.5.8. K. Yosida, An operator theoretical treatment of temporally homogeneous Markov processes, *J. Math. Soc. Japan*, **1** (1949) 244–253.

A.6. *Generalizations*

A.6.1. Yu. N. Blagoveshchenskii, Diffusion processes depending on a small parameter, *Teoriya Veroyatnostei*, **7** (1962) 130–146.

A.6.2. E. B. Dynkin, Markov processes and problems in analysis, *Proc. Int. Cong. Math. Stockholm* (1962) 36–58.

A.6.3. I. I. Gihman, On the theory of differential equations of stochastic processes, *Amer. Math. Soc. trans.*, 2*ème série*. I. (1956) 111–137, 139–161.

A.6.4. K. Yosida, Semi-group theory and the integration problem of diffusion equation, *Int. Cong. Math. Amsterdam*, **1** (1954) 405–420.

B.7. *Statistical Theory of Turbulence*

B.7.1. L. Agostini and G. Bass, *Les théories de la turbulence, Public. Sci. et techn. Ministère de l'Air.*, No. 237, Paris (1950).

B.7.2. G. K. Batchelor, *The theory of homogeneous turbulence*, Cambridge University Press, 1953.

B.7.3. G. Birkhoff and J. Kampé de Fériet, Kinematics of homogeneous turbulence, *J. Math. Mech.*, **7** (1958) 663–704; **11** (1962) 319–340.

B.7.4. F. N. Frenkiel, Etude statistique de la turbulence: fonctions spectrales et coefficients de corrélation, *G. R. A. Rapport technique*, No. 34, Paris (1948).

B.7.5. E. Hopf, Statistical hydrodynamics and functional calculus, *J. Rat. Mech. and Anal.*, **1** (1952) 87–123.

B.7.6. E. Hopf and E. W. Titt, On certain special solutions of the Φ-equations of statistical hydrodynamics, *J. Rat. Mech. and Anal.*, **2** (1953) 587–591.

B.7.7. E. Hopf, Remarks on the functional analytic approach to turbulence, *Proc. Symp. Appl. Math.*, **13** (1962) 157–163.

B.7.8. J. Kampé de Fériet, L'état actuel du problème de la turbulence, *Sci. Aér.*, **3** (1934) 9–34; **4** (1935) 12–52.

B.7.9. J. Kampé de Fériet, Les fonctions aléatoires stationnaires et la théorie statistique de la turbulence homogène, *Ann. Soc. Sc. Bruxelles*, **59** (1939) 145–194.

B.7.10. J. Kampé de Fériet, Fonctions aléatoires et théorie statistique de la turbulence, Chap. 14, pp. 568–628, in A. Blanc-Lapierre and R. Fortet, *Théorie des fonctions aléatoires*, Paris (1953) Masson.

B.7.11. J. Kampé de Fériet, Introduction to the statistical theory of turbulence, *J. Soc. Ind. Appl. Math.*, **2** (1954) 1–9, 143–174, 244–271; **3** (1955) 90–117.

B.7.12. J. Kampé de Fériet, La notion de moyenne dans la théorie de la turbulence, *Rend. Semin. Mat. Fisico Milano*, **27** (1955–1956) 1–43.

B.7.13. J. Kampé de Fériet, Problèmes mathématiques de la théorie de la turbulence homogène, *Cent. Int. Mat. Est. Varenna* (Sept. 1957) Teoria della turbolenza, **1**, 1–104.

B.7.14. R. H. Kraichnan, The closure problem of turbulence theory, *Proc. Symp. Appl. Math.*, **13** (1962) 199–225.

B.7.15. M. Millionstchikov, On the theory of homogeneous isotropic turbulence, *C. R. Acad. Sci. U.S.S.R.*, **32** (1934) 615.

B.7.16. J. Proudman and W. H. Reid, On the decay of a normally distributed homogeneous turbulent velocity field, *Phil. Trans. Roy. Soc. London*, Ser. A, **247** (1954) 163.

B.7.17. T. Tatsumi, The theory of decay process of incompressible isotropic turbulence, *Proc. Roy. Soc. London*, Ser. A, **239** (1957) 16.

B.7.18. G. I. Taylor, Diffusion by continuous movements, *Proc. London Math. Soc.*, Ser. 2, **20** (1922) 196–211.

B.7.19. G. I. Taylor, Statistical theory of turbulence, Parts 1–4, *Proc. Roy. Soc. London*, Ser. A, **151** (1935) 421–478.

B.7.20. N. Wiener, The homogeneous chaos, *Amer. J. Math.*, **40** (1938) 897–936.

B.8. *Burgers Model*

B.8.1. J. Bass, *Les fonctions pseudo-aléatoires, Mem. Sc. Math.*, fasc. CL III (1962).

B.8.2. J. M. Burgers, Application of a model system, to illustrate some points of the statistical theory of turbulence, *Proc. Kon. Ned. Akad. Wetensch.*, **43** (1940) 212.

B.8.3. J. M. Burgers, A mathematical model illustrating the theory of turbulence, *Adv. Appl. Mech.*, **1** (1948) 171–199.

B.8.4. J. M. Burgers, Correlation problems in a one-dimensional model of turbulence, *Proc. Kon. Ned. Akad. Wetensch.*, **53**, Nos. 3, 4, 6 (1950).

B.8.5. J. M. Burgers, On the coalescence of wave-like solutions of a simple nonlinear partial differential equation, *Proc. Kon. Ned. Akad. Wetensch.*, Ser. B, **57**, No. 1 (1954).

B.8.6. J. M. Burgers, Further statistical problems connected with the solution of a simple nonlinear partial differential equation, *Proc. Kon. Ned. Akad. Wetensch.*, Ser. B, **57**, No. 2 (1954).

B.8.7. J. M. Burgers, Statistical problems connected with the solution of a simple nonlinear partial differential equation, *Proc. Kon. Ned. Akad. Wetensch.*, Ser. B, **57**, No. 4 (1954).

B.8.8. E. Hopf, A mathematical example displaying features of turbulence, *Comm. Pure Appl. Math.*, **1** (1948) 303–322.

B.8.9. E. Hopf, The partial differential equation $u_t + uu_x = \nu u_{xx}$, *Comm. Pure Appl. Math.*, **3** (1950) 201–230.

B.8.10. A. Siegel, T. Imamura, and W. Meecham, Wiener-Hermite functional expansion in turbulence with the Bürgers model, *Phys. Fluid*, **6** (1963) 1519–1521.

B.9. *Linear Equations; Well-Set Cauchy Problem*

B.9.1. J. Kampé de Fériet, Sur la mécanique statistique des milieux continus, *Cong. Int. Phil. Sci., Paris* (Oct. 1949); **3** (1959); *Actualités Scientifiques*, 1137, 129–144.

B.9.2. J. Kampé de Fériet, Statistical mechanics of a continuous medium (vibrating string with fixed ends), *Proc. 2nd Symp. Math. Statist. Prob., Berkeley* (1950) 553–566.

B.9.3. J. Kampé de Fériet, Statistical fluid mechanics; two-dimensional linear gravity waves, *Proc. Int. Conf. Part. Diff. Eq.* (June 1960), Madison (1961) Univ. Wisc. Press, 106–136.

B.9.4. J. Kampé de Fériet, Statistical mechanics of continuous media, *Symp. Hydrodynamics Instability;* Amer. Math. Soc., New York (April 1960); *Proc. Symp. Appl. Math.*, **13** (1962) 165–198.

B.9.5. G. Birkhoff and J. Kampé de Fériet, Homogeneous normal random solutions of Cauchy problems (to be published).

B.10. *Banach Spaces; Abstract Cauchy Problem*

B.10.1. G. Birkhoff and T. Mullikin, Regular partial differential equations, *Proc. Am. Math. Soc.*, **9** (1958) 18–25.

B.10.2. G. Birkhoff, Well-set Cauchy problems and C_0-semigroups, *J. Math. Anal. and Appl.*, **8** (1964) 303–324.

B.10.3. E. Hille, The abstract Cauchy problem and Cauchy's problem for parabolic differential equations, *J. Anal. Math.*, **3** (1954) 81–196.
B.10.4. J. Kampé de Fériet, Mesures de probabilité sur les espaces de Banach possédant une base dénombrable, *J. Math. Pure and Appl.*, 9e série, **39**, 97–163.
B.10.5. J. Kampé de Fériet, Les intégrales aléatoires des équations aux dérivées partielles et la Mécanique statistique des milieux continus, *Atti 2° Reunione Math. expression latine, Firenze, Bologna* (1961) 152–166.
B.10.6. J. Kampé de Fériet, Mesures de probabilité sur l'espace de Banach C_0 [0, 1], *C. R. Acad. Sc.*, **255** (1962) 2334–2337.
B.10.7. R. S. Phillips, A note on the abstract Cauchy problem, *Proc. Nat. Acad. Sc.*, **40** (1954) 244–248.

B.11. *Spatial Homogeneity*

B.11.1. J. Kampé de Fériet, L'intégration de l'équation de la chaleur pour des données initiales aléatoires, Jubilé Scientifique de M. Dimitri Riabouchinsky, *Publ. Sci. et Techn. Ministère de l'Air*, Paris (1954) 153–164.
B.11.2. J. Kampé de Fériet, Intégrales aléatoires de l'équation de la chaleur dans une barre infinie, *C. R. Acad. Sc.*, **240** (1955) 710–712.
B.11.3. J. Kampé de Fériet, Random solutions of partial differential equations, *Proc. 3rd Symp. Math. Statist. and Prob., Berkeley*, **3** (1955) 199–208.

B.12. *Conclusion*

B.12.1. J. L. Doob, Semimartingales and subharmonic functions, *Trans. Amer. Math. Soc.*, **77** (1954) 86–121.
B.12.2. J. L. Doob, Martingales and one-dimensional diffusion, *Trans. Amer. Math. Soc.*, **78** (1955) 168–208.
B.12.3. J. L. Doob, A probability approach to the heat equation, *Trans. Amer. Math. Soc.*, **80** (1955) 216–280.
B.12.4. K. Hoffman, *Banach spaces of analytic functions*, Englewood Cliffs, N.J. (1962) Prentice-Hall.
B.12.5. M. Kac, On some connections between probability theory and differential and integral equations, *Proc. 2nd Berkeley Symp. Math. Statist. and Prob.* (1950) 189–216.
B.12.6. J. Kampé de Fériet, Fonctions aléatoires harmoniques dans un demi-plan, *C. R. Acad. Sc.*, **237** (1953) 1632–1634.
B.12.7. J. Kampé de Fériet, Fonctions harmoniques aléatoires dans le cercle unité, *80e Cong. Soc. Savantes, Lille* (June 1955) 411–415.

C. *Appendix*

C.1. J. Kampé de Fériet, Mesures de probabilité sur l'espace de Banach C [0, 1], *C. R. Acad. Sc.*, **245** (1957) 813–816.
C.2. J. Kampé de Fériet, Pseudo-intégrales de Stieltjes aléatoires, *C. R. Acad. Sc.*, **252** (1961) 2162–2165, 3368–3371.
C.3. J. Kampé de Fériet, An elementary theory of the Brownian motion random function, *Sankhya India J. Statist.*, **23**(A) (1961) 207–220.
C.4. N. Wiener, *The Fourier integral and certain of its applications*, Cambridge University Press (1933).

Lectures on Modern Mathematics

In Three Volumes

Edited by T. L. SAATY,
Arms Control and Disarmament Agency,
formerly of the Office of Naval Research.

All those desiring to keep abreast of the major achievements in the various fields of mathematics will welcome this work. Not only will it cover virtually every principal area of modern mathematics but it is written to reach as wide an audience as possible and to make readily accessible a maximum amount of material.

Each of the eminent contributors delineates a substantial research area. He describes it broadly and comprehensively and contributes to this description his individual evaluation of the aesthetic and practical aspects of the field, its position in mathematical development as a whole, and its future, as it is implied in the conjectural exposition of its unsolved problems.

The monographs are expanded versions of expository lectures, conceived and organized by the editor and given at the George Washington University under the sponsorship of the University and the Office of Naval Research. In writing the monographs, the lecturers have successfully compressed the enormous amount of material at their command and at the same time have preserved an informal approach, have refrained from any intricate excursions that might intimidate the novice, and have given each monograph the personal flavor of their own involvement in mathematical research.

JOHN WILEY & SONS

New York • London • Sydney

Saaty

Lectures on Modern Mathematics *Volume III*

510
LEC

Wiley